DOCUMENT RESUME

ED 123 622 CS 002 715

AUTHOR Bormuth, John R., Ed.
TITLE Readability in 1968.
INSTITUTION National Council of Teachers of English, Champaign, Ill.
PUB DATE 68
NOTE 52p.; Originally entered as ED028022

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS *Preschool Children; *Readability; Reading Comprehension; Reading Interests; Reading Level; *Reading Materials; Reading Readiness; *Reading Research

ABSTRACT

This bulletin presents four papers on how to control, manipulate, and predict the readability of printed materials. The first paper describes trends in readability brought about by research tools developed by psychologists and linguists. The second paper explores the effects of word frequency in printed materials on comprehension and concludes that we tend to recognize and understand readily the more frequently used words. Experimental studies which contribute to the improvement of the readability and teaching capacity of preschool books are discussed in the third paper which also illustrates how readable preschool books may be created by exploiting children's love for television and cartoons. The fourth paper is an evaluation of the use of the cloze procedure to determine the suitability of reading materials. Each article is followed by a brief summary and a bibliography. These articles were originally published in the 1967 and 1968 issues of "Elementary English." (NS)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED123622

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Readability in 1968

A Research Bulletin
Prepared by a Committee
of the National Conference
on Research in English

John R. Bormuth, *Editorial Chairman*
University of Chicago

Published for the NCRE by the
National Council of Teachers of English

CS 002 715

The articles in this bulletin
were originally published in
issues of *Elementary English*,
Volumes XLIV and XLV,
December, 1967, and
January, February, March, April, 1968.

The National Conference on Research in English is an organization of one hundred active members qualified to conduct scientific research in English. The purpose of the organization is to stimulate and encourage research in the teaching of English and to publish results of significant investigations and of scientific experimentation. The 1968 president is Albert J. Harris, Bureau of Research, City University of New York.

Copyright 1968
National Conference on Research in English

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

National Conference on Research in English

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

Contents

Introduction v
John R. Bormuth

New Developments in Readability Research 1
John R. Bormuth

The Role of Word Frequency in Readability 7
George R. Klare

Experimental Studies of Readability 18
E. B. Coleman

The Cloze Readability Procedure 40
John R. Bormuth

Introduction

The papers contained in this bulletin address themselves to the question of what makes the language in materials easy or difficult to read. This question is of central importance in the educational process, for much of what a student learns he acquires through his study of written instructional materials. The research which concentrates on finding the answers to this question has traditionally been called readability research.

The chief aim of early readability research was to devise formulas which educators could use to determine if materials were suitable to their students. Because psychological and linguistic theories were as yet poorly developed, the objectives of the early researchers were necessarily limited and pragmatic. That is, they merely sought to devise an easily applied formula which would predict the reading ease of materials. They could make little headway in attempting to find the sources that caused the difficulty.

Modern readability researchers view their task more broadly. They seek to establish scientifically the principles which will permit us not only to predict the difficulty of materials but also to write the materials to have any level of difficulty we choose. Stated another way, instead of just trying to find correlations between the features of language and its difficulty, modern readability researchers seek to analyze the psychological processes involved in language comprehension and composition so that they can understand why those correlations exist. This is not to say, however, that readability researchers have abandoned the use of correlational research techniques. To do so at this stage in the development of our knowledge of readability, would be folly. There are literally thousands of features of language, any number of which might be an important factor in reading ease. As yet, psycholinguistic theory is too crude to permit the *a priori* exclusion of any of these variables as an important factor. Until such a theory has been constructed, researchers must simultaneously consider a large number of variables. Correlational techniques provide the best available method for doing so.

The point that should be emphasized is that, while modern readability researchers retain some of the techniques used by earlier researchers, they have abandoned much of the purely pragmatic approach which pervaded much of the early research. The ultimate goal of modern correlational research is to construct a theory of readability which will permit us to reduce this area of research to an experimental science, for to be sound a technology of written instruction can be based only on the findings of experiments.

Education stands to gain much from the fact that readability researchers have defined their task more broadly. Out of modern readability research is growing a body of scientific knowledge about the nature of language comprehension and language composition. This knowledge is providing the basis out of which a technology of written instructional materials is growing. This knowledge is providing the basis of a theory of the psychological nature of the processes people utilize in the comprehension and composition of language. With this knowledge it will be possible to make substantial improvements in the curriculum and instruction in reading comprehension and language composition.

The papers contained in this series reflect this newer approach to readability. Their authors analyze their subject matters in much greater detail, base their analyses upon the general bodies of linguistic and psychological theory, and address themselves to the question of how readability may be deliberately controlled and manipulated as well as to how it may be predicted.

John R. Bormuth
Editor

JOHN R. BORMUTH

New Developments in Readability Research*

Teachers of reading and language arts are avid students of readability research. The objective of readability research is to find out what features of language influence the difficulty children have in learning the knowledge expressed in language. The results provide the specialists with the information they need to tailor instructional materials to fit the reading abilities of their students and with readability formulas by which they can determine if materials already prepared are suitable for their students. Finally, by studying how the many features of language influence comprehension, readability research provides insights into the nature of the comprehension process, itself.

Recent Advances. The past few years have seen rapid and somewhat startling developments in readability research. For example, the readability formulas available only three years ago could, at best, predict only 25 to 50 percent of the variation we observe in the difficulties of instructional materials. Today, we have not one but several prototype formulas which are able to predict 85 to 95 percent of the variation. This high level of precision represents an improvement of from 35 to 75 percent over the validities of older readability formulas. The purpose of this paper is to describe the nature and results of some of the recent research and the efforts currently being made to forge our newly gained knowledge into practical educational tools.

Areas of Advancement. Among the most important events leading to the present developments was the publication of two books summarizing the readability research done up to that time. One was by Chall (1958) and the other by Klare (1963). From these books it became clear that future readability research had to concentrate on three problems. First, a more reliable method had to be developed for measuring the difficulty children have in understanding materials. Second, researchers had to develop a real sophistication in linguistics so that they could learn to measure and describe the linguistic features of materials that are really important in affecting comprehension. Third, investigators had to analyze their data in far more detail than they had up to that time. What follows is an account of what resulted when efforts were made to attack each of these problems.

Measurement of Comprehension Difficulty

Problem. Until recently, investigators used multiple choice tests to determine the comprehension difficulties of materials. They made a test over each of the passages they were studying, tested the students after they had read each passage, and then found the mean percentage of questions answered correctly. The test means represented the difficulties of the passages.

Dr. Bormuth is an Associate Professor of Education at the University of Chicago.

* This article has been reprinted from the December, 1967, issue of *Elementary English*, Volume XLIV, pages 840-845.

This method presented two problems. First, because the test was itself a reading task, the investigator was never quite certain whether his test mean measured the difficulty of the passage or just the difficulty of the test questions. Second, these tests could tell him nothing about how difficult each word, phrase, or sentence in the passage was.

Construction of Cloze Tests. Shortly before Chall and Klare published their books, Taylor (1953) reported his first work with the cloze procedure. The cloze readability procedure can be used to make tests from any passage of written language. To do so, the investigator selects the passage he wishes to study, deletes every fifth word, and replaces the deleted words with underlined blank spaces of a standard length. The test is given to children who have not previously read the passage, and they are instructed to write in each blank the word they think was deleted. Their responses are scored correct when they exactly match the words deleted, but disregard minor misspellings.

Advantages. Cloze readability procedure does not confuse the measurement of passage difficulty by injecting an extraneous reading task into the process. It also has the added advantage that investigators can measure the difficulty of every word, phrase, or sentence in a passage.

Research. The cloze readability procedure drew the attention of readability researchers who set about studying cloze tests to see if they were valid and reliable measures of the comprehension difficulties of passages. Their research has become too extensive to review here, but Bormuth (1967a) and Rankin (1964) have each published detailed analyses of this research. In general, the research showed that cloze readability tests are highly valid and highly reliable measures of the comprehension abilities of students and of the comprehension difficulties of materials.

Description and Measurement of Language

Early researchers felt a need to make their formulas so simple they could be used even by clerks having little technical knowledge of language. For example, to determine the complexity of a word, the analyst either counted its syllables or looked it up to see if it was on a list of words thought to be easy. To determine the grammatical complexity of a sentence, the analyst counted the number of words, and sometimes the number of prepositions, in the sentence. While it was, at that time, important for formulas to be simple, the old formulas vastly over-simplify the rich array of language features that influence its comprehension difficulty. The over-simplification also contributed to the fact that the old formulas were inaccurate.

Vocabulary Complexity: Present investigators are probing more deeply into the question of what makes a word difficult to understand. It is a gross oversimplification to say that the words on some list have been shown to be easier to understand. This leaves us still asking which of a word's many meanings did children understand and why those words are easier for students. What follows is a discussion of some of the features of words currently being investigated.

Word Length. Children have always thought of long words as hard and short words as easy. Researchers have recently rediscovered this fact and begun investigating word length as a variable. Coleman (1961) found that the mean length of the words in passages has a correlation of -.90 with passage difficulty when length is measured either in terms of letters or syllables. Bormuth (1966) found correlations -.76 and -.68, respectively, for the same

measures. The differences in the sizes of the correlations found by Coleman and Bormuth probably resulted only from differences in the variabilities of the passages they studied.

Morphological Complexity. A word is often a complex structure which may be analyzable into a stem and a series of inflectional, derivational, and lexical affixes. This may be an important source of difficulty in understanding words. Coleman (1966) found that passage difficulty had a correlation of -.88 with the mean number of affixes and stems into which the words in a passage can be analyzed and a correlation of the same size with the number of inflectional morphemes. However, it should be noted that this analysis yields almost the same results as counting the number of syllables in words.

Latin Base Syllables. Many of the words in English contain syllables which can be traced back to a Latin origin. Words containing Latin base syllables give the subjective impression of being abstract. Coleman and Aquino (1967) (7) have found that the difficulties of passages have a correlation of -.81 with the proportion of their syllables that had Latin bases.

Abstractness. Although there are almost as many meanings of the word abstractness as there are people who use it, nearly everyone agrees that, whatever it is, it has an influence on the difficulty of a word. Coleman (1966) devised a definition which permitted him to count reliably the proportion of nouns that referred to internal mental states and found that this number had a correlation of -.78 with passage difficulty.

Frequency. It has long been known that the frequency with which a word is used has some influence on the difficulty people have in understanding it. But, frequency was thought to be a weak variable since Lorge (1949) had found only a correlation of .51 between it and difficulty. More recently, Bormuth (1966) has shown that frequency and difficulty have a curvilinear relationship and that, when this fact is taken into account, they have a correlation of .66. Klare (1967) has taken the position that the frequency of a word may directly reflect most of the other characteristics of the word.

Grammatical Complexity: The degree of intricacy of the grammatical relationships between the parts of a sentence has always been considered an important source of the difficulty people have in understanding the sentence. Until recently, the chief means of assessing grammatical complexity consisted of counting the number of words in sentences. Two major objections can be raised to considering sentence length as the sole factor affecting grammatical complexity. First, it forces us to accept the dubious proposition that all sentences containing the same number of words possess the same degree of complexity. Thus, we are asked to believe that the sentence *The man saw the boy who found the penny which was lost* has the same degree of complexity as *The penny which the boy whom the man saw found was lost.* Second, the number of words in a sentence does not measure a natural unit of language. We cannot simply add or chop off a few words to make the sentence more or less complex. Making a sentence more complex may or may not increase the number of words it contains; and increasing the number of words it contains may or may not increase the complexity of a sentence.

The grammatical complexity of a sentence actually results from the grammatical structure of the sentence. Consequently, modern researchers are investigating measures of grammatical complexity based on the grammatical structures of sentences. Their approach is firmly supported by the

experiments performed by Martin (1966) and Johnson (1966a, 1966b) which demonstrate that people utilize the phrase structure of sentences as they process the sentences.

Syntactic Depth. Yngve (1960) developed an analysis which obtains the number of grammatical facts a reader must temporarily hold in his memory as he reads a sentence. Presumably, the more grammatical facts the reader must hold in his memory at any one time as he reads a sentence, the more likely he is to forget one of those facts and the more likely he is to fail to comprehend some aspect of the sentence. Martin (1966) has shown that people's responses to sentences are closely related to the depth measures of the sentences. Bormuth (1966) found a correlation of -.55 between depth and passage difficulty. Further, he found (1963) that the effects produced by depth were independent of those produced by sentence length.

Modifier Distance. A variation on the depth measure was developed by Bormuth (1967) and is being investigated by him and by Coleman and Aquino (1967). This variable measures the number of words occurring between a word or phrase and the word or phrase it modifies, on the theory that the longer a grammatical fact is held in memory, the more likely it is that it will be forgotten. Preliminary results indicate that there is a correlation of -.80 and -.90 between this feature and passage difficulty.

Transformational Complexity. A sentence such as *The little boy ran* may be represented as resulting from a transformation which embedded the kernel sentence *The boy was little* into the kernel sentence *The boy ran.* Chomsky (1965) has argued that to interpret a sentence people must transform a sentence back into its kernel sentences.

An interesting aspect of the transformation analysis is the fact that it can be used to measure what early researchers referred to subjectively as being the idea density of materials. Coleman (1966) found that the proportion of words that were derived by nominalizing verbs and adjectives had correlations of -.76 and -.57, respectively, with passage difficulty. Many parts of speech represent transformations, also. Bormuth (1966) found that counts of the various parts of speech exhibited correlations as high as .81 with passage difficulty. For his current studies, Bormuth (1967b) (1) has developed an inventory of what seems to be all the transformations found in English and is studying the effects of each transformation on difficulty.

Contextual Variables: Modern researchers are looking beyond the word and the sentence to find the features of language that operate over longer segments of text to influence comprehension. Rosenberg (1966) found indications that passages containing words which people tend to associate with each other are easier to recall. Coleman and Aquino (1967) are finding that anaphoric analyses yield variables that predict passage difficulty. Anaphora are words or phrases which refer back to an earlier word or phrase in a passage. The use of anaphora indicates the extent to which a passage deals in depth with a single topic. Since the work in this area is only beginning, it is still too early to predict its outcomes, but, it seems likely that gains in this area will have great value in increasing our ability to predict and control passage difficulty.

Readability Formulas

Early investigators had to defer the investigation of many important problems until research in other disciplines had made tools available for studying those problems.

As may be seen from the preceding discussion, linguistic research provided readability researchers with new and powerful tools for analyzing language. Similarly, research tools became available for studying the problems involved in designing readability formulas. As a result, we have now learned enough to design much sounder readability formulas.

Readability and Reading Ability: A problem long troubling researchers was the question of whether the features that influenced readability for poor readers also influenced the readability of materials for more able readers. If the same features of language influence readability for both and by the same amount, then a single and fairly simple formula can be used to predict readability for all students, regardless of their level of accomplishment in reading. But if different features influence difficulty for students of differing levels of reading achievement or if the same features influence difficulty by different amounts, then we must develop more complex and materially different kinds of formulas. Bormuth (1966) studied this problem and found that, regardless of the person's reading ability, the same features of language that caused difficulty for him caused the same amount of difficulty for others.

This fact was doubly interesting because some people held that only teachers of young children had to concern themselves with the readability of instructional materials. It can now be said that language has just as strong an influence on what adults are able to learn from materials as it does on what children are able to learn.

Shapes of the Relationships: A second question was whether or not the relationship between language variables and the difficulty of that language was linear. For example, is the difference in difficulty between two and three syllable words as great as the difference in difficulty between seven and eight syllable words? If not, the simple correlation techniques used by early researchers yield misleading results. Bormuth (1966) found that many of the relationships showed varying degrees of curvature. For example, adding another syllable to a one syllable word increases its difficulty far more than adding another syllable to a seven syllable word. The same is true of many other features. Interestingly, it was the variables most frequently used in the old formulas that showed the greatest amount of curvature. Hence, future readability formulas must include appropriate transformations of measurements taken of these features.

Form of the Formulas: The old readability formulas were presented in the form of what is called a multiple variable, linear equation. These equations have a characteristic that makes them unsuitable for use as readability prediction formulas. To use them we must assume that any correlation observed between two variables, say sentence length and word length, must always exist and that it must be of the magnitude observed in the original research. This simply is not true of the language features used in most formulas. For example, it is a simple matter to write long or short sentences using words of any length we choose. The result is that the old formulas yield misleading results whenever the correlation is anything other than the correlation the formulas assume. Most future readability formulas will probably be designed to provide a profile of the level of difficulty represented by each of the language features in a passage.

Summary

Readability researchers have made rapid strides in the past few years, increasing

the accuracy of readability formulas by as much as 75 percent. The reason lies largely in the fact that researchers in several disciplines have developed research tools which have aided greatly the study of readability. Psychologists have developed the cloze procedure into an accurate and reliable method of measuring language difficulty. Linguists have developed descriptions of various features of language and these descriptive devices have been further adapted into powerful new techniques for measuring the features of language that influence its comprehension difficulty. Finally, advances in our understanding of the mathematics used in our analyses have led to improved designs for readability formulas. The result of these advances is that, within a year or two, educators will have placed in their hands powerful new tools for determining if instructional materials are suitable for use with their students. Of greater long range importance, we will gain much more insight into the comprehension processes and into the processes by which language may be made more understandable.

References

1. Bormuth, J. R., *Implications of Cloze Procedure for the Evaluation of Instructional Programs*. Los Angeles: University of California, Center for the Study of Evaluation of Instructional Programs, Occasional Report No. 3, 1967.
2. Bormuth, J. R., *Development of Readability Analyses*. Project in Progress at the University of Chicago, Sponsored by the U. S. Office of Education, 1967a.
3. Bormuth, J. R., "Readability: A New Approach," *Reading Research Quarterly*, 1 (1966b) 79-132.
4. Chall, Jeanne S., "Readability: An Appraisal of Research and Application," *Ohio State University Educational Research Monograph* No. 30, 1958.
5. Chomsky, N., *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: M.I.T. Press, 1965.
6. Coleman, E. B., "Developing a Technology of Written Instruction: Some Determiners of Complexity of Prose," Symposium on Verbal Learning Research and the Technology of Written Instruction. 1966. (Unpublished Manuscript)
7. Coleman, E. B. and M. R. Aquino, *Communications Project*. Project in Progress at the South West Regional Laboratory in Los Angeles. 1967.
8. Johnson, N. F., "The Influence of Associations Between Elements of Structured Verbal Responses," *Journal of Verbal Learning and Verbal Behavior*, 5 (1966a) 361-368.
9. Johnson, N. F., "On the Relationship Between Sentence Structure and the Latency in Generating the Sentence," *Journal of Verbal Learning and Verbal Behavior*, 5 (1966b) 369-374.
10. Klare, G. R., *The Measurement of Readability*. Ames, Iowa: Iowa State University Press, 1963.
11. Klare, G. R., "The Role of Word Frequency in Readability," 1967. (in press)
12. Lorge, I., "Readability Formulas—An Evaluation," *Elementary English*, 36 (1949) 85-95.
13. Martin, E. and K. H. Roberts, "Grammatical Factors in Sentence Retention," *Journal of Verbal Learning and Verbal Behavior*, 5 (1966) 211-218.
14. Rankin, E. F., "Cloze Procedure—A Survey of Research," *Yearbook of the South West Reading Conference*, 14 (1965) 133-148.
15. Rosenberg, S., "Recall of Sentences as a Function of Syntactic and Associative Habit," *Journal of Verbal Learning and Verbal Behavior*, 5 (1966) 392-396.
16. Taylor, W. L., "Cloze Procedure: A New Tool for Measuring Readability," *Journalism Quarterly*, 30 (1953) 415-433.
17. Yngve, V., "A Model and an Hypothesis for Language Structure," *Proceedings of the American Philosophical Society*, 104 (1960) 444-466.

GEORGE R. KLARE

The Role of Word Frequency in Readability*

Over the years many hundreds of language variables have been found to be predictively related to readability. These have even included such seemingly unlikely ones as percentage of words beginning with *w*, *h*, *b*, *i*, and *e*. Furthermore, those mentioned in the recent analyses of Bormuth (3) and Coleman (10) indicate that the number of such variables is probably limited more by the linguist's resourcefulness in quantifying new characteristics of language than by anything else. One who wishes to understand and control readability is therefore faced with the problem of making sense of this profusion of variables.

One way to avoid having to deal with such an unmanageable number is to combine the several most highly predictive. This is what the developers of readability formulas have done. The common procedure, of course, has been to start with the most predictive variable, then add the next most predictive, *etc.*, in a regression equation. Though formulas with eight or more variables have been put together in this way, most formulas have been limited to two or three variables. The reason is simple enough: beyond this point the amount of predictive accuracy added becomes very small compared with the amount of additional work needed for formula application.

Word difficulty and sentence difficulty

A look at the two variables most often used in formulas indicates that they are (1) some aspect of word difficulty, usually frequency, and (2) some aspect of sentence difficulty, usually length. Of the 31 formulas published up to 1960 [see Klare (31)], 17 use a word-count factor directly and most others a related factor (*e.g.*, word length). Similarly, 12 use a sentence-length factor directly and many others use it indirectly (*e.g.*, proportion of simple sentences in the sample sentences). Two factor analyses of the Gray and Leary matrix (21) of readability variables indicate essentially the same two factors to be the most important. Brinton and Danielson (5) found that a "vocabulary" factor accounted for the greatest variance and a "sentence" factor the next greatest. Stulorow and Newman (51) found that "relative difficulty of words" accounted for 34 percent and "relative sentence difficulty" 20 percent of the total variance of 93 percent contributed by the top ten factors they found. (None of the remaining ten accounted for more than 8 percent of the variance, and no others were therefore interpreted.)

No attempt will be made here to examine in detail both word and sentence factors. Attention will be restricted to words, and in particular to word frequency, not only because it has been more predictive but

Dr. Klare is a Professor of Psychology and Dean of the College of Arts and Sciences at Ohio University, Athens.

* This article has been reprinted from the January, 1968, issue of *Elementary English*, Volume XLV, pages 12-22.

also because studies of its effect provide a better understanding of readability. Included implicitly with frequency, however, will be the sub-concept of "familiarity," as determined, for example by Dale (12) and used in his formula with Chall. As shown by Noble (44), rated familiarity is determined almost uniquely by frequency (to the extent of an index of relationship of .998 in his work).

Some values of word frequency

This first importance of word frequency is not so surprising when viewed in the light of language usage. It has long been known that humans do not use different words equally often, even in the long run. Instead, some words are repeated very frequently and others used very infrequently. The extent of the repetition, however, is much greater than is usually realized. For example, French, Carter, and Koenig (15) have reported that the 100 most frequent words made up 75 percent of the total of approximately 80,000 words in 500 telephone conversations. Coleman (11) has found that in speech the 100 most frequent words make up approximately 49 percent of the total. It has been found that though speech contains somewhat more repetition than writing, the same tendency does indeed hold. Wendell Johnson (29) has collected a large body of data on this entire phenomenon; following the philosopher Charles Peirce, he has also recommended the use of "types" for the number of different words in a passage and of "tokens" for the total number of words One of Johnson's students, Mary Mann, found that 25 percent of the total of 67,200 tokens used by 24 university freshmen in writing their life stories consisted of only 10 types (*the, I, and, to, was, my, in, of, a,* and *it*).

G. K. Zipf (60) has, in fact, found the same general trend for many different language situations, as has E. L. Thorndike (55), among others. Zipf has referred to this phenomenon as "equilibrium," and has attempted to quantify it in the "standard curve" of English words. He has also attempted to explain it in terms of what he called the "Principle of Least Effort," which he saw operating in human behavior other than language. The principle refers, briefly, to minimizing the "probable average rate of work" required to reach a goal (*i.e.*, send a particular message). Others (*e.g.*, Mandelbrot) have suggested modifications of Zipf's work, but the basic notion of humans minimizing the energy they use when communicating remains a central feature of language usage and one of the presumed bases for frequency findings.

A closely related aspect also deserves comment: the tendency for words to become shorter as they are used more frequently. Zipf is usually credited with first having shown this relationship, arriving at it from the observation that words become shorter with length of time in the language (time and usage being themselves correlated). Dramatic examples can be seen in the reduction of "television" to "video" to "TV" (or "television" to "telly" in Great Britain); "horseless carriage" to "car"; or the reductions of agency names to letter abbreviations such as NRA or WPA in the alphabet days of the Thirties.

What Zipf showed was that this process occurred, in somewhat less dramatic fashion perhaps, with most words. In the process he gave this frequency-shortening principle its clearest and most detailed statement. Actually, like many another scientific observation, it was made earlier and simply went unnoticed by most writers. Mata V. Bear (1), in an unpublished Master's thesis eight years earlier at the University of Chicago, found a close correlation between word length and frequency of usage. Furthermore, this prin-

ciple found early application in readability formulas. Pear had found the percentage of monosyllabic words in a selection to provide a fair index of reading difficulty, and G. R. Johnson (25) made this the basis of his formula.

Frequency and readability

Perhaps the earliest recorded evidence that frequency was related to readability, however, came over a thousand years earlier. Lorge (36) pointed out that the Talmudists at that time made word and idea counts of their manuscripts so that they could use frequency of occurrence to distinguish usual from unusual meanings. Recognition of frequency's importance is corroborated in a general way by Gray (20), who indicates that study of the elements of difficulty began centuries ago in connection with children's reading; concrete evidence of this appeared in 1840 when ease of understanding in the McGuffey Readers was considered in terms of vocabulary.

Important in the early work was the list of 1,500 familiar Russian words provided by N. A. Rubakin in 1889. Just afterward in 1898, the German F. W. Kaeding published his mammoth word count. It provided a scientific base for the relation of vocabulary to reading difficulty and established a basic vocabulary foundation (see Lorge). It also set the stage for the 1921 appearance of E. L. Thorndike's *The Teacher's Word Book* (57). This book was of special importance because it became one of the bases not only for the teaching of vocabulary in the schools but also for the first measure of readability that can really be considered a formula, that of Lively and Pressey (34) in 1923. Two similar word books were subsequently published by Thorndike (56), and all three played an important role in the developmental history of formulas.

Word frequency as a variable

It is not surprising, then, to find word frequency appearing as a variable in some way in most attempts at readability measurement. What *is* surprising is that relatively little attention has been paid to a closer examination of *how* word frequency affects readability. It appears possible, however, to gain at least a somewhat better understanding by examining the many recent experiments which have yielded behavioral correlates of word frequency. The remainder of this paper will be concerned with such studies. No attempt has been made to provide a complete review of: vocabulary studies, [a good reference here is Dale and Razik (13)]; word lists and frequency counts [see Fries and Traver (16) and Bongers (2)]; or even all studies providing behavioral correlates of frequency [Brown (6) reviews studies relating frequency to recognition threshold up to about the end of 1959, and Goss (19) covers studies relating frequency to paired associate learning].[1]

Attention, then, has been focussed on those studies that can provide some help in understanding how and why word frequency is related to readability. Before this question can be meaningfully examined, however, a clear picture is needed of just what is meant by "readability." A review of the many validity studies of readability formulas [see Klare (31)] indicates that the readability of a passage or text can be operationally defined in terms of the following.

1. *Efficiency of reading.* The two variables most often used to measure ef-

[1] A list of 35 additional references which are relevant to the remainder of this discussion is available from the author. Note also the appearance of Davis Howes', "A Word Count of Spoken English," *Journal of Verbal Learning and Verbal Behavior*, 5 (December, 1966) 572-606.

ficiency are average number words read per minute (or per second) and average number of words read per visual fixation (or its converse, average number of fixations per word).

2. *Reader judgment.* This is sometimes measured directly by asking readers to rate or rank from two to a large number of passages in terms of readability or preference; sometimes it is measured indirectly by determining the readership of two or more alternative versions of a passage or article.
3. *Comprehension and learning.* This is most often measured by means of comprehension tests of the typical multiple-choice type, but the "cloze" procedure has come to be used recently. [This latter method usually requires that subjects fill in correctly the blank spaces substituted for every fifth word of a passage; see Taylor (53).] In addition to comprehension tests, measures of amount learned (or immediate retention) and amount retained (or delayed retention) have also been used.

Frequency and reading efficiency

With this background in mind, examination of the behavioral correlates of frequency as related to readability can be undertaken. First, the reading efficiency aspect of readability. A relevant series of studies is that concerned with the effects of word frequency upon tachistoscopic measurement of recognition and report, which is very similar to measurement of reading speed. These were precipitated by a controversy over the relative ease of recognition of neutral versus emotionally disturbing words.

Historically, the experiment that seems to have initiated the chain of studies was that of Bruner and Postman (7). These authors found that words of an emotionally disturbing nature were more difficult to recognize in tachistoscopic presentation than neutral words. They came to describe the raised recognition threshold as "perceptual defense." A related finding was that some emotionally disturbing words were less difficult to recognize than others, which they described as "perceptual sensitization." The latter term seemed to disappear from the literature, but the former became the focus for a prolonged controversy having effects still being felt.

In the years immediately following, the study of changes in recognition threshold centered about the effect of personal values and value-related words [see Postman, Bruner, and McGinnies (48)]. The arguments concerning perceptual defense gathered force, however, with publication of a study by McGinnies (40) of socially disapproved, or "taboo," words. In his work he noted that such words had significantly higher recognition thresholds than neutral words, and suggested that the difference was due to the fact they were socially taboo. McGinnies, however, did not control for differences in word frequency ratings. The taboo words used were subsequently shown by Howes and Solomon (24) to have had much lower frequency ratings than the neutral words. The threshold difference, they suggested, may well have been due to this and, perhaps, to response suppression. McGinnies (39) subsequently denied this, on the basis of his data, and the controversy over the existence of perceptual defense and its causes began in earnest. The nature of perceptual defense itself is not of concern here, but the behavioral correlates of frequency thus initiated are. This latter topic will therefore be the focus of attention.

Familiarity and recognition

The probable relationship of familiarity of words to recognition and reaction were

not completely unknown prior to this time. Bruner and Postman, in their earlier study, indicated that the familiarity to subjects of some of the words they used *may* have been a factor along with perceptual defense. But about this time Solomon and Howes were engaged in studies which *clearly* demonstrated the relationship of frequency of occurrence of words in English usage to recognition threshold. In one study [Howes and Solomon (23)], 75 words differing in Thorndike-Lorge frequency count were presented tachistoscopically. It was found that visual duration threshold was an approximately linear function of the logarithm of the relative frequency with which a word occurs in print. The authors found product-moment correlations between the two variables ranging from —.68 to —.75, the negative values indicating that high frequency words have a low threshold and low frequency words a high threshold. Many subsequent investigators have verified this close inverse relationship between word-count frequency and threshold.

Use of an estimate of language frequency such as the Thorndike-Lorge tables provide cannot, of course, accurately indicate the frequency with which a given subject (reader) will have seen a given word. Solomon and Postman (50) therefore used experimentally controlled frequencies of exposures (occurrences) ranging from 1 to 25 with pronounceable nonsense words as stimuli. Recognition thresholds once again varied inversely with frequency of prior usage. Many other investigators reported similar results. The better control made possible by using experimentally produced frequency rather than the less precise estimate based on samples of language, such as the Thorndike-Lorge tables provide, can be seen in a study by King-Ellison and Jenkins (30). These authors reported the extreme correlation of —.99 between tachistoscopic exposure time and the logarithm of frequency.

These studies clearly suggest that word frequency, by its effect on visual recognition threshold, can increase reading efficiency through more rapid word recognition. Exactly *how* this effect on recognition threshold is produced, however, became the subject of a new controversy. It was first clearly stated in the work of Goldiamond and Hawkins (18). The presumption in the studies up to that point was generally that increased frequency aided *visual perception*. Goldiamond and Hawkins, however, demonstrated that it affected *response probability* or response bias. They used nonsense syllables with frequencies built up experimentally in the manner of Solomon. Instead of presenting these syllables in the tachistoscope for recognition, however, they presented no verbal stimuli at all. Their subjects were not aware of this, having been told that the stimuli were present but below threshold, and when asked to respond they naturally gave nonsense syllables. The relationship between the logarithm of frequency of presentation and "recognition threshold," however, was of the same order as that found when stimuli *had* been presented, as in the studies cited above.

Goldiamond and Hawkins' work initiated a new set of studies designed to resolve the visual perception versus response probability (or bias) explanations. Conflicting results were found, some favoring a perceptual and some a response explanation, and some both. Perhaps the best conclusion that can be drawn is that of Zajonc and Nieuwenhuyse (58): response bias can be clearly demonstrated when no stimuli are present, but when they are present and when a stringent recognition criterion is employed, response bias plays a negligible role. Put another way, it might be said that frequency of occurrence of words can have

both effects, depending upon the "reading" circumstances.[2] Each of these might be seen in the reading efficiency aspect of readability, as discussed below.

Readability as efficiency

Cherry (8) has suggested that a reader hypothesizes what he will see in an immediately forthcoming fixation on the basis of what he has seen in the immediately preceding fixations. Thus, frequency of occurrence of groups of words in such phrases as "of the . . ." or "by means of the . . ." would provide an occasion for response bias to occur. That is, when a reader sees the first word of such a phrase, he is likely to expect the second and would doubtless respond with it if asked to hypothesize what it might be. In fact, it has been shown in studies of cloze procedure and of "predictability" [Rubenstein and Aborn (49)] that such response bias can indeed occur. In the former studies, subjects were (usually) asked to fill in every fifth word deleted from a text (as indicated earlier); in the latter, subjects were asked to predict each successive word of a text. The high probability of correct response that was found is doubtless a function of many factors, such as syntactic rules and familiarity with the topic, but frequency also clearly plays a part. While it is uncommon for a reader actually to find words deleted from his text (as in these studies), it is not so uncommon for the similar situation in which his attention wanders or in which a word is written or printed unclearly. It is in such instances that response bias, as a correlate of frequency, may contribute to readability.

Frequency may also contribute to readability when words are present and are clearly seen by the reader. The way in which this effect may occur is related to the early finding (historically) that words are perceived as patterns. It seems quite clear that, as Bricker and Chapanis (4) have stated, recognition of complex stimulus patterns such as words is not all-or-none. As exposure values increase, partial perceptions of increasing degrees of completeness and accuracy occur, and eventually a subject reports a stimulus correctly.

The effect of frequency in this situation can be seen in a study by Haseley (22). He tested Klare's hypothesis that the amount (percentage) of a verbal stimulus necessary for recognition would be a function of its frequency of occurrence. He used a "word mutilator" with a mask to successively expose portions of English words of differing Thorndike-Lorge frequencies and words of differing experimentally-created frequencies. He found a relationship between the logarithm of frequency and the percentage of a word necessary for recognition that was especially clear-cut in the case of the nonsense words.

Word frequency has another related effect on reading efficiency. As indicated earlier, it has been shown that words tend to become shortened with increased time and usage in a language. This in turn affects efficiency through the fact that shorter words have a lower recognition threshold than longer words [see McGinnies, Comer, and Lacey (41) or Newbigging and Hay (43)].

Visual fixations and readability

This tendency appears to play a part in the finding of Klare, Shuford, and Nichols

[2] After this paper was written an article by D. E. Broadbent, "Word-frequency Effect and Response Bias," appeared [*Psychological Review*, 74 (January, 1967) 1-15]. In it, the author argues that a response bias hypothesis is sufficient to explain the results of the available studies. He believes that common (frequent) words are perceived more readily than uncommon because of a prior bias in favor of the common words, which combines with sensory evidence favoring the objectively correct word. As can be seen, this is more a re-definition of "response bias" than a disagreement with the two-effect explanation suggested here.

(32) that the measures of number of words read per visual fixation and number of words read per second are both significantly increased in more readable as opposed to less readable material. A subsequent study (unpublished) was made to determine whether or not the above effects were due *entirely* to the fact that the average word length was shorter in the more readable than the less readable version. In this study, added white space was placed between the words in the more readable version so that it occupied the same total linear space as the less readable version. It was found that while the number of words read per fixation no longer differed significantly, the number of words read per second still showed a significant difference favoring the more readable version. Thus the fact that frequency of usage tends to produce shortened words can be seen to play a part, at least insofar as the fixation measure of efficiency is concerned. But as the reading speed measure shows, this is not the only part played by frequency; of two words of equal length, the more frequent will still tend to be reported first.

To summarize the effects of frequency of occurrence of words upon the reading efficiency aspect of readability, the following may be said. The frequency of occurrence of the words used is probably a major factor in affecting both words read per fixation and per second in textual material. As frequency is increased, increases in reading efficiency are produced. The limits of this effect are not known, although studies of frequency and of practice effect suggest it is probably asymptotic, with the greatest effect produced by increasing the frequency of low-frequency rather than high-frequency words. Nor is it clear whether the result is due to response bias or perceptual facilitation, although both appear likely under certain reading circumstances. Frequency plays a further role through its relationship with word length, since shorter words have a lower threshold than longer (and since less of a frequent word need be seen for recognition to take place).

Frequency and preference

Since word frequency clearly appears to play a part in the reading efficiency aspect of readability, does it also play a part in the acceptability of (preference for) more readable as opposed to less readable material? Relevant data here are much less numerous than in the case of reading efficiency. A number of studies have shown that readers prefer a more readable to a less readable version of a passage, but few studies suggest the extent to which this may be due to word frequency. Studies by Swanson (52) and Ludwig (38), however, showed greater readership for text versions with shorter or easier words versus versions with longer or harder words. A study by Klare, Mabry, and Gustafson (33) showing a close relationship between judgments of versions as "easy to read" and as "pleasant to read" (tetrachoric correlations ranging from .86 to .97) also suggests that such preferences may be based upon word frequency (through its effect upon reading efficiency).

A series of studies by Ronald Johnson and his associates lends some support to this hypothesis a well as suggests a more direct relationship between frequency and preference. Johnson, Thompson, and Frincke (28) found that greater frequency of occurrence of English words was related to increased "goodness" on the semantic differential (to the extent of a correlation of .63). Rated goodness, in turn, was related to visual duration threshold (and thus, presumably, to reading efficiency). These authors found, further, that manipulation of the frequency of nonsense

words produced a systematic variation in their rated goodness. Frincke and Johnson (17) also found that the relation between frequency and goodness persisted even when pronunciability was held constant through the use of homophones. In subsequent studies, Johnson, Frincke, and Martin (27) and Johnson, Weiss, and Zelhart (26) found goodness related to meaningfulness, and in the first study once more found goodness related to visual duration threshold. Newbigging (42) similarly found "bad" words or the semantic differential to have higher thresholds than "good" or "neutral" words.

In summary, the effects of frequency of occurrence of words upon preference for a more readable over a less readable version appears to result from one or the other of two possible effects, or from both. First, the frequency of occurrence of a word is related to its rated goodness, thus producing a possible direct effect upon preference. Second, the rated goodness of a word is related to its visual duration threshold, thus producing a possible indirect effect upon preference by affecting the reading ease of a text, since ease and preference are themselves related.

Frequency and comprehension

The evidence thus indicates that frequency of occurrence of words affects both the reading ease and preference aspect of readability. Does this variable also affect the comprehension and learning aspect of readability? As for comprehension, it was precisely because educational experience indicated that more common words were more comprehensible that the first word frequency studies were made. As a corollary to, and refinement of, their counts of word frequency, Lorge and Thorndike (37) extended this work to counts of the frequency of word meanings (dictionary meanings). Once again, the purpose was to help teachers and textbook writers avoid undesirable rarities of word meanings (and thus increase comprehensibility) in their writing and speaking.

Lorge and Thorndike began their count of word meanings by excluding the 500 most frequent words. They soon discovered, however, that the most common words have the largest variety of meanings, and therefore included them; in fact, their count of these words has been separately published and is perhaps its most useful outcome (35).

Both Zipf (59) and Thorndike (54) have pointed out that the number of dictionary meanings of a word is related to its frequency of occurrence and to its age in the language (presumably an index of frequency). It is logical to ask at this point whether or not the existence of a large number of meanings might not cause interference, thus rendering the more frequent words *less* comprehensible. As this writer has pointed out elsewhere, a hierarchy of frequencies exists among the various dictionary meanings of a word in the semantic count. Thus, use of a highly frequent word will most often tend to involve use of a highly frequent, and presumably comprehensible, meaning.

Frequency and learning

On the question of whether word frequency will clearly affect the *learning* of written material in much the same way it affects comprehensibility there is somewhat contradictory information. Noble has made a series of studies of the relationship of word frequency, familiarity, and meaning. He has measured meaning both as number of synonyms subjects can give to a word (45), which is similar to the Lorge-Thorndike-Zipf definition, and as number of continued written associations subjects can give within 60 seconds (47). In both cases, the relationship between frequency

and meaning indices has been found to be very high (up to an index of relationship of .92). Using his indices, Noble has found that both familiarity of words and word meaning (46) enhance performance in serial verbal learning.

Deese (14), on the other hand, has questioned whether there is any essential relationship between word frequency and the measure of free recall. He believes that free recall is, instead, a function of the readily available associations among high frequency words. Considering the demonstrated relationship between frequency and number of associations, however, suggests that frequency may still be the important underlying factor. Cofer (9), in commenting on Deese's position, takes the point of view that frequency of experience plays a role in recall through "response availability." This term refers to the hierarchical arrangement of words in a person's response repertoire (as "response bias," it was also felt to be a significant factor in the word recognition studies described earlier).

Summary: frequency and comprehension

To summarize, the effects of frequency of occurrence of words upon superior comprehension in more as opposed to less readable material takes several forms. First, increased frequency itself seems to play a role, as does the corresponding increase in available meanings as frequency increases. The existence of a hierarchy among meaning frequencies apparently reduces the interference effects that might otherwise result when numbers of meanings are possible. Second, serial verbal learning appears to be improved, and perhaps free recall and other measures of learning also. This latter may be due either to the presence of a hierarchy of available words or to the increased associations (greater meaning) of high frequency as opposed to low frequency words.

As indicated earlier, the results of studies relating word frequency to comprehension and learning have not been as clearcut as those relating word frequency to reading ease and preference. A possible resolution is suggested by some recent work by Coleman (10). He found a high positive correlation between the frequency of content words in 12 passages and comprehension as measured by summed cloze scores, as would be expected. For function words (such as "the" or "of"), however, he found a high negative correlation between frequency and cloze scores. As he shows, the overall effect of frequency may therefore be lost under certain circumstances. Of particular note is the likelihood that this will occur in speech and in beginning reading. Coleman's work thus suggests that varied and refined counts of frequency may be very fruitful in the study of readability.

Frequency of occurrence of words, as this paper indicates, clearly plays an all-pervasive role in language usage. Not only do humans tend to use some words much more often than others, they recognize more frequent words more rapidly than less frequent, prefer them, and understand and learn them more readily. It is not surprising, therefore, that this variable has such a central role in the measurement of readability. As further knowledge is gained of the circumstances under which word frequency for an individual increases, it is probable that understanding and control of readability can be still further improved.

References

1. Bear, Mata V., *The Length of Word as an Index of Difficulty in Silent Reading*. Unpublished Master's Thesis, University of Chicago, 1927.
2. Bongers, H., *The History and Principles of Vocabulary Control*, Parts I, II, and III. Holland: Wocopi-Woerden, 1947.
3. Bormuth, John R., "Readability: A New Approach," *Reading Research Quarterly*, Spring, 1966.
4. Bricker, P. D. and Alphonse Chapanis, "Do

Incorrectly Perceived Stimuli Convey Some Information?" *Psychological Review*, 60 (1953) 181-188.

5. Brinton, J. E. and W. A. Danielson, "A Factor Analysis of Language Elements Affecting Readability," *Journalism Quarterly*, 35 (Fall, 1958) 420-426.
6. Brown, William P., "Conceptions of Perceptual Defense," Monograph Supplements, *The British Journal of Psychology*, 1961.
7. Bruner, Jerome S. and Leo Postman, "Emotional Selectivity in Perception and Reaction," *Journal of Personality*, 16 (1947) 69-77.
8. Cherry, Colin, *On Human Communication*. New York: John Wiley and Sons, Inc., 1957, pp. 286-288.
9. Cofer, Charles N., "Comments on Professor Deese's Paper," in Chapter 2 of C. N. Cofer (Ed.), *Verbal Learning and Verbal Behavior*. New York: McGraw-Hill Book Co., 1961.
10. Coleman, Edmund B., "Developing a Technology of Written Instruction: Some Determiners of the Complexity of Prose," Symposium on Verbal Learning Research and the Technology of Written Instruction, Columbia University, 1966. Unpublished.
11. Coleman, Edmund B., "Words Rank Ordered as to Frequency," 1966. Unpublished.
12. Dale, Edgar, "Familiarity of 8000 Common Words to Pupils in the Fourth, Sixth, and Eighth Grades." Columbus, Ohio: Bureau of Educational Research, Ohio State University. Mimeographed.
13. Dale, Edgar and Taher Razik, *Bibliography of Vocabulary Studies* (Second Revised Edition). Columbus, Ohio: Bureau of Educational Research and Service, the Ohio State University, 1963.
14. Deese, James, "Frequency of Usage and Number of Words in Free Recall: The Role of Association," *Psychological Reports*, 7 (1960) 337-344; also, "From the Isolated Verbal Unit to Connected Discourse," Chapter 2 in C. N. Cofer (Ed.), *Verbal Learning and Verbal Behavior*. New York: McGraw-Hill Book Co., 1961.
15. French, H. R., C. W. Carter, Jr., and W. Koenig, Jr., "The Words and Sounds of Telephone Conversation," *Bell System Technical Journal*, 9 (1930) 290-324.
16. Fries, Charles C. and Aileen A. Traver, *English Word Lists*. Washington, D. C.: American Council on Education, 1940.
17. Frincke, Gerald and Ronald C. Johnson, "Word Goodness and Word Frequency in Homophone Pairs," *Psychological Reports*, 7 (1960) 470.
18. Goldiamond, Israel and William F. Hawkins, "Vexierversuche: The Log Relationship Between Word-Frequency and Recognition Obtained in the Absence of Stimulus Words," *Journal of Experimental Psychology*, 56 (1958) 457-463.
19. Goss, Albert E. and C. F. Nodine, *Paired-Associates Learning: The Role of Meaningfulness, Similarity, and Familiarization*. New York: Academic Press, Inc., 1965.
20. Gray, William S., "Progress in the Study of Readability," in *Library Trends*. Chicago: University of Chicago Press, 1937, pp. 237-254.
21. Gray, William S. and Bernice E. Leary, *What Makes a Book Readable*. Chicago: University of Chicago Press, 1935.
22. Haseley, Leonard, *The Relationship Between Cue-Value of Words and Their Frequency of Prior Occurrence*. Unpublished Master's Thesis, Ohio University, 1957.
23. Howes, Davis and Richard S. Solomon, "Visual Duration Threshold as a Function of Word-Probability," *Journal of Experimental Psychology*, 41 (1951) 401-410.
24. Howes, Davis and Richard S. Solomon, "A Note on McGinnies' 'Emotionality and Perceptual Defense,'" *Psychological Review*, 57 (1950) 229-234.
25. Johnson, George R., "An Objective Method of Determining Reading Difficulty," *Journal of Educational Research*, 21 (April, 1930) 283-287.
26. Johnson, Ronald C., R. L. Weiss, and P. F. Zelhart, "Similarities and Differences Between Normal and Psychotic Subjects in Responses to Verbal Stimuli," *Journal of Abnormal and Social Psychology*, 68 (1964) 221-226.
27. Johnson, Ronald C., Gerald Frincke, and Lea Martin, "Meaningfulness, Frequency, and Affective Character of Words as Related to Visual Duration Threshold," *Canadian Journal of Psychology*, 15 (1961) 199-204.
28. Johnson, Ronald C., Calvin W. Thompson, and Gerald Frincke, "Word Values, Word Frequency, and Visual Duration Thresholds," *Psychological Review*, 67 (1960) 332-342.
29. Johnson, Wendell, *People in Quandaries*. New York: Harper and Brothers, 1946, Appendix.
30. King-Ellison, Patricia and James J. Jenkins, "The Durational Theshold of Visual Recognition as a Function of Word-Frequency," *American Journal of Psychology*, 67 (1954) 700-703.
31. Klare, George R., *The Measurement of Readability*. Ames, Iowa: The Iowa State University Press, 1963, pp. 75-80.
32. Klare, George R., Emir H. Shuford, and William H. Nichols, "The Relationship of Style Difficulty, Practice, and Ability to Efficiency of Reading and to Retention," *Journal of Applied Psychology*, 41 (1957) 222-226.
33. Klare, George R., James E. Mabry, and Levarl M. Gustafson, "The Relationship of Style Difficulty to Immediate Retention and to Acceptability to Technical Material," *Journal of*

Educational Psychology, 46 (May, 1955) 287-295.

34. Lively, Bertha A. and Sidney L. Pressey, "A Method of Measuring the 'Vocabulary Burden' of Textbooks," *Educational Administration and Supervision*, 9 (October, 1923) 389-398.
35. Lorge, Irving, *The Semantic Count of the 570 Commonest English Words*. New York: Bureau of Publications Teachers College, Columbia University, 1949.
36. Lorge, Irving, "Word Lists as Background for Communication," *Teachers College Record*, 45 (May, 1944) 543-552.
37. Lorge, Irving and Edward L. Thorndike, *A Semantic Count of English Words*, Vols. 1 and 2. New York: The Institute of Educational Research, Teachers College, Columbia University, 1938. Microfilm.
38. Ludwig, Merritt C., "Hard Words and Human Interest: Their Effects on Readership," *Journalism Quarterly*, 26 (June, 1949) 167-171.
39. McGinnies, Elliott, "Discussion of Howes and Solomon's Note on 'Emotionality and Perceptual Defense,'" *Psychological Review*, 57 (1950) 235-240.
40. McGinnies, Elliott, "Emotionality and Perceptual Defense," *Psychological Review*, 56 (1949) 244-251.
41. McGinnies, Elliott, Patrick B. Comer, and Oliver L. Lacey, "Visual-recognition Thresholds as a Function of Word Length and Word Frequency," *Journal of Experimental Psychology*, 44 (1952) 65-69.
42. Newbigging, P. L., "The Perceptual Redintegration of Words Which Differ in Connotative Meaning," *Canadian Journal of Psychology*, 15 (1961) 133-142.
43. Newbigging, P. L. and Janet M. Hay, "The Practice Effect in Recognition Threshold Determinations as a Function of Word Frequency and Length," *Canadian Journal of Psychology*, 16 (1962) 177-184.
44. Noble, Clyde E., "The Familiarity-Frequency Relationship," *Journal of Experimental Psychology*, 47 (1954) 13-16; also, "Supplementary Report: Familiarity and Frequency," *Journal of Experimental Psychology*, 59 (1960) 432-433, and 60 (1961) 418.
45. Noble, Clyde E., "The Meaning-Familiarity Relationship," *Psychological Review*, 60 (1953) 89-98.
46. Noble, Clyde E., "The Effect of Familiarization upon Serial Verbal Learning," *Journal of Experimental Psychology*, 49 (1955) 333-338; also, "The Role of Stimulus Meaning *(m)* in Serial Verbal Learning," *Journal of Experimental Psychology*, 43 (1952) 437-446.
47. Noble, Clyde E., "An Analysis of Meaning," *Psychological Review*, 59 (1952) 421-430.
48. Postman, Leo, Jerome S. Bruner, and Elliott McGinnies, "Personal Values as Selective Factors in Perception," *Journal of Abnormal and Social Psychology*, 43 (1948) 142-154.
49. Rubenstein, Herbert and Murray Aborn, "Learning, Prediction, and Readability," *Journal of Applied Psychology*, 42 (February, 1958) 28-32.
50. Solomon, Richard L. and Leo Postman, "Frequency of Usage as a Determinant of Recognition Thresholds for Words," *Journal of Experimental Psychology*, 43 (1952) 195-210.
51. Stolurow, Lawrence M. and James R. Newman, "A Factorial Analysis of Objective Features of Printed Language Presumably Related to Reading Difficulty," *Journal of Educational Research*, 52 (March, 1959) 243-251.
52. Swanson, Charles E. and Harland G. Fox, "Validity of Readability Formulas," *Journal of Applied Psychology*, 37 (April, 1953) 114-118.
53. Taylor, Wilson L., "Cloze Procedure: A New Tool for Measuring Readability," *Journalism Quarterly*, 30 (Fall, 1953) 415-433.
54. Thorndike, Edward L., "On the Frequency of Semantic Changes in Modern English," *Journal of Genetic Psychology*, 39 (1948) 23-37.
55. Thorndike, Edward L., "On the Number of Words of Any Given Frequency of Use," *Psychological Record*, 1 (1937) 399-406.
56. Thorndike, Edward L., *A Teacher's Word Book of 20,000 Words* and (with Lorge) *The Teacher's Word Book of 30,000 Words*. New York: Bureau of Publications, Teachers College, Columbia University, 1932 and 1944.
57. Thorndike, Edward L., *The Teacher's Word Book*. New York: Teachers College, Columbia University, 1921.
58. Zajonc, Robert B. and B. Nieuwenhuyse, "Relationship Between Word Frequency and Recognition: Perceptual Process or Response Bias?" *Journal of Experimental Psychology*, 67 (1964) 276-285.
59. Zipf, George Kingsley, "The Meaning-Frequency Relationship of Words," *The Journal of General Psychology*, 33 (1954) 251-256.
60. Zipf, George Kingsley, *"The Psycho-Biology of Language*. Boston: Houghton Mifflin Co., 1935.

E. B. Coleman

Experimental Studies of Readability*

Part I

Stimulus Dimensions that Affect Readability

This paper has two major theses:

1. A promising research strategy for gaining an understanding of language behavior is to collect a huge matrix of functional relations that plot the effects upon language skills exerted by language characteristics. Experimental studies of readability will provide many of these functions.
2. The few studies we have at present, crude and incomplete as they are, will be sufficient to make massive improvements in the teaching capacity of textbooks—at least at the first-grade and preschool level. To make equally massive improvements at the adult level, however, we must study the effect of language characteristics that exist between sentences and between paragraphs. This will require developing efficient techniques for measuring understanding.

Let us say that an experimental study of readability investigates the effect that a characteristic of prose exerts upon some reading response such as speed of reading or degree of understanding. Let us call the characteristic of prose a stimulus dimension and call the measure of readability a response measure. Thus, an experimental study of readability plots a reading response as a function of a stimulus dimension of prose; it plots an S-R function. Linguistics and the psychology of verbal learning provide a list of stimulus dimensions: word familiarity, word length, type-token ratio, abstract noun ratio, clause-to-kernel ratio, *etc.* We also have an extensive list of response measures: the reader's ability to answer questions about the passage, his ability to memorize the passage word for word, his ability to fill in a cloze test based upon the passage, his ability to follow instructions prescribed in the passage, the information he gains by reading the passage, and others.

Thus, we already have a vast taxonomy of stimulus dimensions and an extensive set of response measures. What we need to provide next is a huge matrix of functions —graphs that plot each of the response measures as a function of each of the stimulus dimensions that affects it. This paper will review a restricted sample of the available literature within an outline that will describe how we can provide such a matrix.

In this paper, I will introduce several variables such as spelling and phonic regularity that are usually not included in studies of readability. While these variables are of slight importance in determining

---

Professor Coleman is Chairman of the Psychology Department at the University of Texas in El Paso. At the time this paper was written he was on leave and serving as Project Manager for the Southwest Regional Laboratory in Inglewood, California.

* This article has been reprinted from issues of *Elementary English*. Part I appeared in February, 1968, Volume XLV, pages 166-178; Part II, March, 1968, Volume XLV, pages 316-320, 333.

readability for adults, they are of considerable importance in determining readability for six year olds. A thesis of this paper is that the most fertile field for experimental studies of readability is one that focuses on first-grade material.

In fact, one purpose of this paper is to stimulate interest in experimental studies of readability by showing the massive improvement they could make in the teaching capacity of primary reading materials. I hope to show that given a data base provided by experimental studies of readability, stories can be constructed that are simple enough for a five year old to read after a few minutes of casual instruction—simple enough to enable us to teach preschoolers to read at home by television—simple enough to provide material within the competence of the mentally and culturally deprived—simple enough to make it worthwhile to teach deaf children to read at very young ages and reduce the mental retardation that results from their insulation from language. In short, I hope to show that little books can be constructed which are so easy to read that they will change the whole nature of the readiness program for preschoolers (see Fig. II for an example of such a book).

The characteristics of an effective program of research upon readability will be discussed under three headings:

1. The set of stimulus dimensions should be organized systematically so that we can gain a coherent overview with a minimum number of experiments.
2. There should be an equally systematic array of response measures, including those of most interest such as comprehension and information gain.
3. The experiments should be performed upon relevant populations—relevant populations of language materials as well as relevant populations of readers—and the experimental design should allow us to generalize simultaneously to all populations of interest.

I will discuss the first heading in Part I. The second and third will be discussed in Part II.

1. A Systematic Organization of Stimulus Dimensions

Linguistics and the psychology of verbal learning have provided a large number of stimulus dimensions that affect the difficulty of printed language. We also have a large array of response measures. Obviously we cannot collect all possible S-R functions, only a tiny sample of strategically located ones. Years could be wasted by an unsystematic attack on a problem of this magnitude. A systematic approach would first organize the stimulus dimensions into a grid. Then the investigator could select a few stimulus dimensions spotted at equal steps across the grid and relate them to measures of verbal behavior. If this produced no coherent overview, then additional experiments would relate intermediate stimulus dimensions to the same measures of verbal behavior. As the grid became dotted with systematically located S-R functions, a coherent picture would emerge and the gaps could be bridged by interpolations.

To illustrate a systematic organization of stimulus dimensions, I will use a research program whose purpose is to collect the data that will permit us to design the most readable books possible for first graders.

Above and beyond the magnitude of the improvements that are possible with beginning material, there are important methodological reasons for beginning experimental studies of readability at this level: (1) In beginning reading, the skills can

be analyzed into paired associate learning and the simpler forms of concept formation. Experimental psychologists know how to measure these skills in tightly controlled laboratory experiments. In later reading, however, comprehension, information gain, critical evaluation, esthetic enjoyment, and other skills become important. We know very little about measuring these latter skills; much ancillary work must be done before we can measure them with adequate precision. (2) In beginning reading, the language population is small and well defined; it is restricted to a few hundred common words. In later reading, however, the language population becomes large and poorly defined.

Let us organize the stimulus dimensions that affect the readability of beginning material into some sort of systematic matrix. At the present preliminary state of data collection there is little reason to prefer one organization to another. Here is one: (a) The reader distinguishes the letters. (b) He packages letters and their sounds into words. He also recognizes words as wholes. (c) He packages words into higher grammatical units such as phrases and clauses. (d) He packages these into still higher order units—paragraphs, stories, *etc.*

These four steps will serve to step off our matrix of stimulus dimensions in sufficient detail for preliminary data collection. Later, as data accumulate, the matrix will be graduated in finer detail by inserting intermediate steps.

A. Distinguishing the Letters

It is clear that the readability of material for first graders can be improved by increasing the legibility of the letters. We are not free to design an entirely new set of characters; however, there are many type fonts in use, and surely letters from different fonts can be combined to form an alphabet whose characters are more distinguishable from one another than any existing set.

A recent study that is relevant to the legibility of beginning material is one by Popp (18). Using non-reading kindergarteners as subjects, Popp compared each letter with every other letter. Her result was a confusion matrix showing how often these children confused each letter with every other letter. Rank-ordered as to the number of times they were confused with one another, the most frequently confused pairs were: *p-q, d-b, d-q, d-p, b-p, h-u, i-l, k-y, t-u, c-e, d-h, h-n, h-y, j-k, n-u.*

Popp presented her letters in isolation, but in reading, the letters are imbedded in words and the child is able to distinguish upward protruding from downward protruding letters. It is not unreasonable to assume that the most frequently confused pairs might be different if the letters were presented to the children embedded in pseudowords. In an unpublished experiment, Himelstein (12) tested this assumption, but her pattern of results was almost the same as Popp's. Himelstein tachistoscopically projected a stimulus letter surrounded by two *o*'s (*opo*) and then asked the child to pick out that letter from a pair surrounded by *o*'s (*opqo*). She tested only the pairs that Popp had found most frequently confused. Rank-ordered as to number of times they were confused by the children, Himelstein's pairs were: *p-q, d-b, d-p, b-p, t-u, b-q, i-l, k-y, h-n, d-h, h-u, c-e, h-y, n-u, j-k.*

By combining letters from different type fonts and by making slight changes in the characters, it is easy to make many of the above pairs very distinguishable from one another. Some letters may be printed in small versions of the capital shapes so as to eliminate confusions between *h*-H, *u*-U, *L*-*i*, *etc.* The letter *q* is always followed by *u*. Why not print them as a ligature (*qu*) and thus eliminate confusions between *q* and *p*.

If these suggestions were followed, of course the letters could be designed to be more esthetically pleasing than in their above form.

I have tested modifications similar to the above in a series of experiments. My most recent experiments compared traditional to experimental versions of the following pairs: *p-q, d-b, d-p, b-p, h-n, u-n,* and *i-l*. A letter was projected tachistoscopically for .10 seconds between two *o*'s (*odo*), and a non-reading kindergartener was asked to pick that letter from a pair surrounded by *o*'s (*odbo*). As might be expected, every experimental pair was significantly more legible than its conventional counterpart (by at least the .05 level). The surprising finding was the amount of difficulty the children had in distinguishing the conventional *p-q* and *d-b*. Our preschoolers did no better than chance—78 errors to 74 correct choices.

To summarize the experimental studies of legibility as they affect beginning readers: the conventional alphabet contains a number of pairs that are frequently confused by beginning readers. It would be a simple matter to alter the pairs slightly so as to make them less confusing. Furthermore, since most of these alterations would simply require that we select letters from several type fonts (such as capitals and lower case), almost no negative transfer should occur. We would not be adding any new characters to be learned; we would simply be changing their sequence of introduction for young children.

B. Packaging Sounds and Letters into Words

This step is concerned with look-and-say reading and with phonics. There are experimental studies of readability that suggest ways to improve readability according to either subskill of reading.

Look-and-Say Reading. Woodworth (24) summarized a group of studies that suggested that word shapes are recognized as wholes. The studies suggested that words would become easier to distinguish from one another if the alphabet were redesigned to increase the distinctiveness of word shapes. Specifically, the studies suggested that word shapes would become more distinctive if we created an alphabet with equal numbers of upward protruding, downward protruding, and non-protruding letters—also with an equal number of angular and curved letters. Since 1938, there have been a number of sporadic tests of Woodworth's suggestion (*e.g.*, Byrne, 3), but no sustained program of research on the problem.

Recently, in an unpublished study, Desberg (10) showed that word shape can play an important role in word recognition by non-reading preschoolers. Ten trios of pseudowords that differed to a minimum degree in word shape were prepared (e.g., *ocen, uosa, suam*) and compared to trios that differed to a maximum degree (e.g., *qhur, ruyc, coan*). A pseudoword was projected tachistoscopically and ten non-reading preschoolers were asked to select it from a trio. All ten children made fewer mistakes when selecting words whose shapes differed to a maximum degree ($p < .001$). The experiment was interpreted as evidence that a lower-case alphabet could be constructed that would increase the legibility of words for beginning readers. Such an alphabet can be constructed by substituting a few capitals for their lower case counterparts and by making a few other equally slight changes.

To summarize the studies of an alphabet that would maximize distinctiveness of whole word shapes: the studies to date have not shown word shape to be a large source of variance in determining readability for adults; nevertheless, there is reason to believe that a sustained program of

research could develop an alphabet that would significantly improve readability for first graders.

Length. The length of words clearly affects the ease of recognizing them in look-and-say learning. Most readability formulas have some measure of word length, but length in syllables, length in morphemes, and length in letters are all highly correlated. They can, however, be disentangled. To the extent that words are processed visually, length in letters should affect readability. To the extent that words are processed vocally, length in syllables should affect readability. To the extent that they are processed mentally, length in morphemes should affect readability.

In an incomplete, factorial design, I recently measured the independent effects of length in letters, length in syllables, and length in morphemes. I tachistoscopically projected six words for 1.0 second and asked my subjects (college students) to memorize the six. My measure was number of exposures to memorization There was no significant effect due to syllables. Figure I shows the effect due to letters and morphemes. To the extent that memory for a list of words is part of reading, length in letters (visual processing) and length in morphemes (mental processing) appear more important than length in syllables (vocal processing).

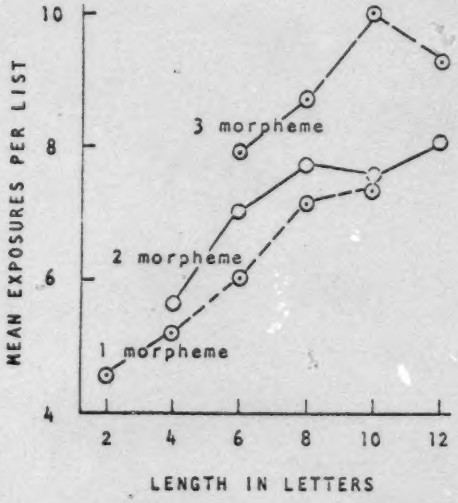

Figure I
Mean exposures to memorization plotted as a function of word length in letters. The parameter is word length in morphemes.

Frequency. The frequency with which a word is used affects the ease of recognizing it. Klare (14) has provided a convenient summary of the effects of frequency on readability.

It need only be added that the usual word counts may be deceptive for first graders. For first graders, the frequency counts should be based on speech, not printed material. There are certain words that occur frequently in print (*will, would, did, am,* etc.), but seldom as whole words in speech. These words usually occur in reduced form (*'ll, 'd, 'd, 'm,* etc.), and they may be difficult for the child to recognize.

Phonics: Learning the Sound-Letter Associations. All readers sound out words or parts of words, but this reading skill has more influence on readability in beginning reading than in that of later years. Before a child can sound out a word, he must have learned the sound-letter associations. This is ordinary paired associate learning, and according to the usual three-stage analysis (Underwood and Shultz, 22), the degree to which a child learns any particular sound-to-letter association would depend upon three factors: the legibility of the letter, the response strength of the sound, and any previous association between the two.

We have already discussed the legibility of letters in A above. There are several indices available that measure the degree to which the child has mastered the phonemes of English. One measure would be studies such as those of Templin (21) that measure the age at which the child first uses the

phoneme. Carterette and Jones (4) have provided a second measure. They tallied the frequency with which the different phonemes occurred in children's speech. They found that the consonant phonemes occurred in the following order when ranked from highest to lowest: *n, r, t, m, s, d, w, l, k,* etc.

In an ordinary paired associate learning task, Berdiansky (1) measured the ease with which non-reading preschoolers learned a number of sound-to-letter associations, that is, he measured the ease with which they learned to pronounce certain letters. As might be expected, he found a strong frequency effect: the child learned the pronunciation to be associated with a certain letter very easily if the pronunciation was a phoneme what occurred frequently in English.

Berdiansky's experiment suggests that for beginning readers at least, a story might be more readable if the pronunciation of its letters were frequently occurring phonemes. Before discounting phoneme control as impractical, one should remember that the vocabulary of beginning material is tightly controlled. Just as we control vocabulary, we can also control phonemes. A gifted cartoonist can write a story with only three or four words. If the letters and phonemes as well as the words were carefully chosen, the story could be read by an average four year old (see Fig. II).

This is not to say that Berdiansky's experiment provides us with all the data we need for phoneme selection. Among other things, we need to know if a child will induce the basic phonic concept (sounding out new words) more easily if his beginning reading material is restricted to certain classes of phonemes. For instance, sounding out an unfamiliar word would seem more difficult if the beginning material consisted mostly of stops and affricates (*p, b, t, d, k, g, j,* and *ch*) because these sounds are most difficult to pronounce in isolation. The child inevitably learns sounds such as *puh, buh,* etc. If Berdiansky's experiments were supplemented by a dozen or so similar ones, we would be able to practice phoneme selection in beginning reading materials as we now practice vocabulary restriction.

Phonics: Regularity. One of the subskills of reading is the ability to sound out words. For first graders in particular, it is reasonable to assume that readability of material will vary according to the degree of phonic regularity in that material and also according to the particular phonemes permitted in that material. Currently there are two major systems for regularizing English, the Initial Teaching Alphabet (ITA) and Bloomfield's linguistic system. Both have considerable flaws.

The Initial Teaching Alphabet must misspell 75% of the words in English, and about 40% of these changes are radical ones. There are several reasons for this, but the most profound one is that the pronunciation of many of the English morphemes changes if the morphemes change environment.[1] Consider the different pronunciations of the past tense morpheme in trott*ed*, pass*ed*, and cross*ed*. Consider the following derivations: *democratic-democ-*

[1] More generally, when one fits an alphabet to a language, there are three factors that must be considered. First is letter-to-sound regularity, phonic regularity. But there are two others of almost equal importance. Second is morpheme regularity; morphemes should be spelled the same way in different environments (trott*ed*, pass*ed*, cross*ed*, *sign*, *sign*ificant, *etc.*). Third—and this becomes a critical factor in a worldwide language such as English—people that speak widely differing dialects should be able to read the same language; a Kentucky hillbilly and a Cockney dockworker should be able to read the same books. One must compromise between these three factors in fitting an alphabet to a language. Clearly, the problem is much more difficult than designing a character for each phoneme of English; it requires wisdom plus a considerable linguistic competence.

racy, expedite-expedient, sign-significant, etc.

In short, the misspellings of the Initial Teaching Alphabet guarantee a considerable amount of negative transfer when the child is shifted to traditional orthography. An extensive series of tightly controlled laboratory experiments on transfer have shown conclusively that negative transfer occurs any time an organism learns a response to a stimulus that he must later unlearn in order to substitute another response to that stimulus. Of course, the Initial Teaching Alphabet has proved to be an excellent teaching technique so it also guarantees a considerable amount of positive transfer, and the weight of evidence suggests that the positive transfer more than outweighs the negative transfer. Nevertheless, we should reduce the negative transfer as much as possible; we could reduce it considerably if we marry the Initial Teaching Alphabet to Bloomfield's linguistic system.

In addition to its theoretical contributions, Bloomfield's linguistic system (2) has shown that it is possible to select a large number of English words that are spelled regularly, that is, that are spelled with internal consistency. But his linguistic system suffers from one fatal flaw; many of the most important words in English are not spelled regularly (*was, were, would, should,* etc.). Thus, if one restricts himself to Bloomfield's list of words, he restricts his reading materials to such sentences as "*Fat Pat sat on the mat.*" The most phlegmatic child can quickly be reduced to hysterics if he is forced to read too much of this.

Thus, we have two systems for regularizing English, each with great advantages; each with great disadvantages. Fortunately, it is possible to combine them in proportions that maximize the advantages of each and minimize the disadvantages. One simply spells a few dozen absolutely necessary function words in an Initial Teaching Alphabet and selects the remainder of his words from Bloomfield's list. Thus, negative transfer is reduced to an absolute minimum; only a few dozen words are misspelled in the transitional alphabet. Furthermore, it is possible to select those few dozen words so that their misspelled shape in the transitional alphabet will be almost identical to their shape in the traditional orthography. Thus, by combining the Initial Teaching Alphabet and Bloomfield's linguistic list in the proper proportions, we can get almost all of the advantages of each and almost none of their disadvantages. Furthermore, if we restrict ourselves to the most common words in Bloomfield's list, we can get almost all of the advantages of the basal readers and almost none of their disadvantages.

Phonics: Phoneme Selection. We noted that insofar as reading involves phonic skills —sounding out words—readability would be affected by phonic regularity and by the particular phonemes used in the materials. We have discussed phonic regularity. Let us discuss improving readability by restricting the phonemes used in the materials.

Except for Berdiansky's experiment mentioned earlier, there is almost no evidence that tells us which letter-sound associations are easier to learn according to any reading subskill. If we are to make beginning reading materials more readable by phoneme selection, we must guess from a very inadequate data base. Except for the frequency counts of Carterette and Jones, about the only information we have are two rough observations: (1) It is almost impossible to make an isolated sound that resembles the stops and affricates (*p, b, t, d, k, g, j,* and *ch*). (2) On the other hand, many of the continuents can be closely approximated in isolation, and in addition, they play a meaningful role in the lan-

guage (*s-s-s* is the sound of a snake or of gas escaping, *m-m-m* is the sound we make when we see something good to eat, *ee-ee-ee* is the sound we make when we are

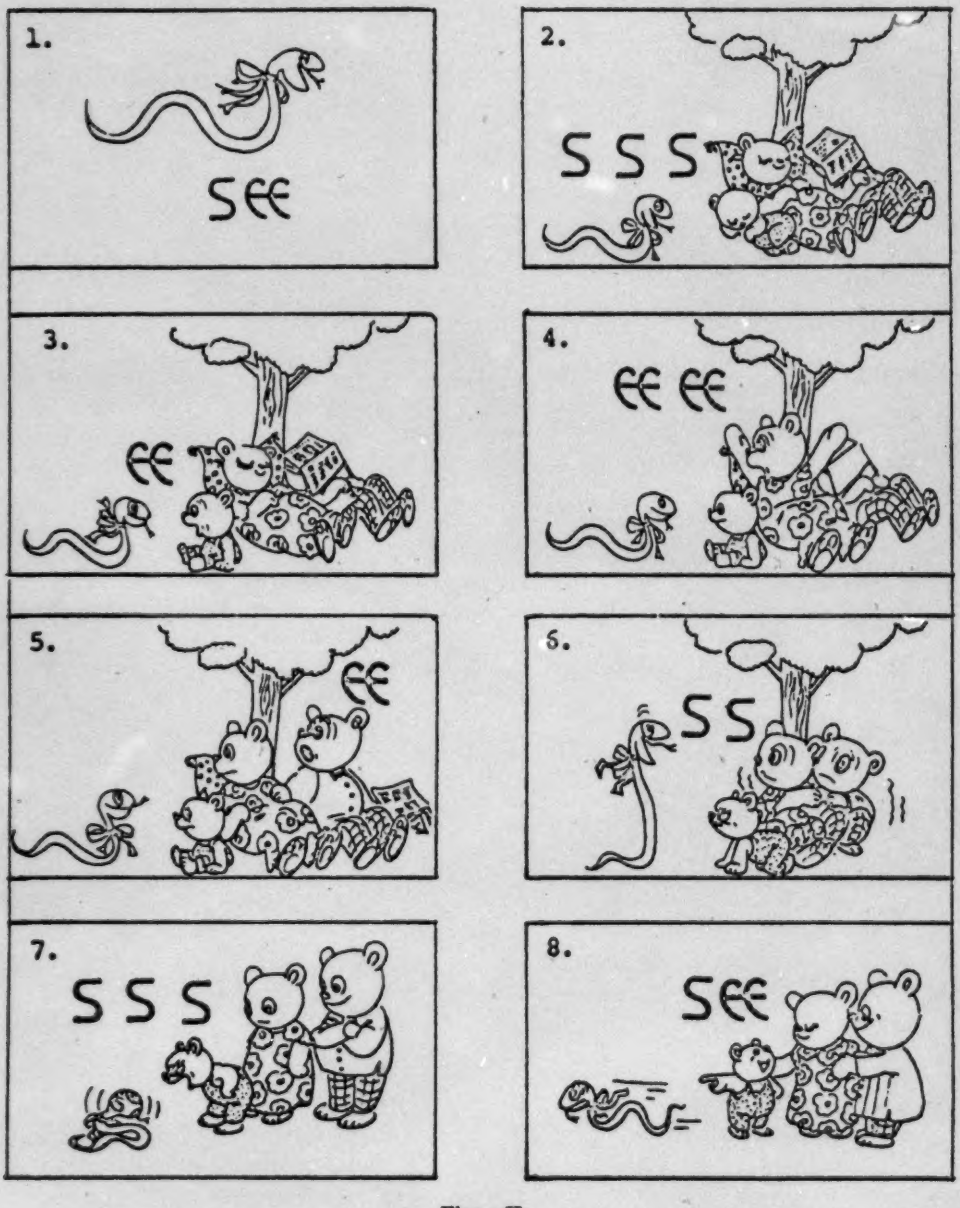

Figure II

Excerpts from a preschool reader that illustrates a technique that can be used for teaching phonics to five year olds when the component sounds of words play a meaningful role in the story.

frightened, *etc.*). Based on these two observations (to the extent that reading involves the phonic subskill), beginning reading materials might be more readable if their phonemes were restricted to the continuents. Then the first words in the beginning material would be ones such that their component sounds could play a meaningful role in the story (*see, am, at,* etc.). This allows a number of techniques for teaching phonics and spelling that are not available if the beginning materials include such words as *come* and *look* (see Fig. II).

Phonics: Blending. Most first-grade teachers would agree that phonic blending is one of the most difficult skills to teach. Blending is less difficult if the first words are composed of sounds in which the consonant and vowels are pronounced with the tongue and other articulators in roughly the same position; the blending transition from one to the other is relatively slight; there is little difference between the two sounds pronounced in isolation and the two blended into a word. Sounds such as /s, sh, n, ee, a, i/ form words such as *an* and *see* that are easiest to blend. Compare the blending transition from vowel to consonant in *she* versus that in *out* by pronouncing them slowly and comparing the relative motion of tongue and lips.

Let us summarize the important points of this section: reading is a complex, interlocking hierarchy of skills, and materials that are most readable according to one subskill are very unreadable according to other subskills. It is possible, however, by simultaneously considering all the subskills when one selects his materials to get a set of materials that approaches maximum readability according to all subskills. We can get the advantages of phonic regularity by using Bloomfield's list of regularly spelled words. We can get the advantages of meaningful, interesting sentences by supplementing that with a very small number of function words misspelled in a transitional alphabet (*thee, iz, wuz,* etc.). Furthermore, we can gain most of the advantages of the basal readers—the effects of frequency—by restricting our selection of regularly spelled words to those that occur most frequently in English. Perhaps—although this must be verified by additional experiments—one can gain still further advantages if he restricts the phonemes in his beginning reading materials to certain continuents and to words whose component sounds can play a meaningful role in his story.

An example may be helpful. In selecting the most readable set of materials for a beginning reading program being developed at Southwest Regional Laboratory (see the example), we used the following procedure: We punched into a computer all the regularly spelled one-, two-, and three-phoneme words of English. We also punched in each word's frequency of occurrence according to its Rinsland count for first graders (19). We assumed that our first materials would be most readable if the phonemes were restricted to the following phonemes: */s, m, ee, I, a, i, n, th, sh,* and *t/*. Then we asked the computer, "Which phoneme should we introduce next? Which phoneme will combine with the present phonemes to form a maximum number of common words?" The letter *w* can be combined with these letters to add a maximum number of common words, and this is the letter that should be used in the next book. Adding *w* to our letters, we repeated the question to the computer. Continuing these steps gave a sequence that enables our books to illustrate the concepts underlying spelling and phonics with a minimum of instruction in letter-sound associations; each letter-sound association introduced in a book gives the child the maximum payoff in useful new words.

The language characteristics considered

thus far are most important in determining readability for beginners; the ones that follow are important for adults also.

C. Packaging Words into Phrases, Clauses, and Sentences.

The greatest number of experimental studies of readability lie under this subheading. Let us subdivide them according to Hockett (13) into those concerned with content morphemes and those concerned with the framework of function morphemes. That is, in the following sentences the content morphemes are in lower case letters and the framework of function morphemes is in capitals:

THE man operatED THE boat skillfulLY.

THE man'S operaTION OF THE boat WAS skillful.

Other sets of content words that can fit in the same framework of function morphemes are boy-explain-problem-quick; team-excavate-tunnel-speedy.

Content Words. In a sustained series of studies, Deese (9) has shown that words are organized in the mind into highly systematic structures. That is, words and sets of words are associated with one another and each word is more tightly associated with some words than with others. Words are organized into word classes, grammatical classes, and semantic classes.

Deese's students have shown that lists of words that are correlated with this mental organization are much easier to learn and memorize than lists that are not (*e.g.*, Miller, 17, Weingartner, 23). Similarly, if the content words of a sentence correlate with this mental organization, the sentence is easier to learn or read (Coleman, 7). In this experiment, noun-is-adjective sentences (*The thimble is bright*) whose noun and adjective had strong pre-experimental associations were compared to sentences whose noun and adjective had weaker pre-experimental associative strength (*The thimble is shiny*). The direction of the results was not surprising; the sentences whose content words had high strength associations were much easier to learn. The magnitude of the results, however, was somewhat surprising. After a single exposure, the subjects would remember 14.5 responses out of 24 paired associates if there were strong pre-experimental associations between noun and adjective. If there were no pre-experimental associations between the noun and adjective, they could remember only 1.58 responses; there was a ratio of 9 to 1 in relative learnability.

To summarize, we can make tremendous gains in readability if we design our materials to correlate with the way the words are organized mentally. This could be accomplished by using cloze tests. That is, if the majority of subjects in a cloze test disagree with the original wording, it would suggest that the wording should be changed. Giving cloze tests is an expensive operation, but the expense might be justified for materials that are used very frequently such as elementary reading material.

Framework of Function Morphemes. Many studies can fit under this subheading, and most of them can be described as comparisons of grammatical transformations. These experiments are concerned with a number of the factors that are usually included when prescribing rules for writing readable material. In fact, most of the time when a writer applies rules for improving readability, he is actually choosing one grammatical transformation above another. Thus, the rules for writing readable material could be stated more precisely in terms of grammatical transformations.

The reader of this paper can verify this for himself by noting the number of times he makes grammatical transformations as he revises the following sentence to make it more comprehensible (*An inclusion of this is an admission of its importance*). If

he tries to make this sentence more readable by increasing the number of personal words, decreasing clause length, or by applying almost any of the rules for improving readability, he will discover that the change he frequently makes is a grammatical transformation—a grammatical transformation that operates upon entire clauses and alters many of these variables simultaneously. The point is this: when we consider the actual operations we perform to improve readability, it seems that grammatical transformations are fundamental units and that describing some rules of readability in terms of smaller units is more or less artificial.

To save space, a sample of the experimental studies that seem most closely related to readability will be summarized in a table.

Table I

Sample of Experiments that Studied Grammatical Relations

Reference Number	*Grammatical Relation*	*Response*	*Finding*
5	Nominalization vs Active Verb	Cloze Score	Active verb was most readable.
6	Nominalization vs Active Verb	Recall	Active verb was most readable.
8	Active vs Passive	Recall	Active was most readable.
11	Active vs Passive	Verification	Active was most readable.
15	Depth	Recall	Sentences of low depth were most readable.
8	Embeddedness	Recall	Readability increased as embeddedness decreased.
20	Order in a N is N Sentence	Recall	Term at end is better remembered.

The prescriptions for readable writing that are given by the above experiments can be summarized briefly; prefer grammatical transformations that give short clauses and use active verbs; prefer grammatical transformations that do not use abstract nouns nominalized from verbs.

In the experiments of Table I, note that grammatical transformations that contain abstract nouns are less readable than their counterparts containing active verbs. The lesson to be gained from this is that abstractness is not necessarily an inalterable, inherent stimulus dimension. As often as not, complexity due to abstractness is superfluous complexity; as often as not, the number of abstract nouns in a passage can be reduced without changing the content of the passage. When a passage contains a large number of abstract nouns nominalized from verbs (*operation*), the writer can easily transform them to active verbs (*operate*). For example, *The exclusion of that candidate is emphasis on the importance of our group* includes three abstract nouns. Let us transform it to a version that

includes none: *When they excluded that candidate, it emphasized that our group was important*. Clearly the abstractness of the first version was not due to inherent abstractness of the subject matter.

In short, there is prose that is abstract for no better reason than that the writer chose one derivative of the verb instead of another.

Note that the above discussion of abstract nouns nominalized from verbs is related to the advice that readable writing should contain a large number of verbs. Note also that this does not necessarily mean that the writer should invigorate his prose with verbs like *excoriate* and *fulminate*. The proportion of verbs can be increased by transforming abstract nouns nominalized from verbs to active verbs. Perhaps another example will be worthwhile. Consider the sentence from the second paragraph of this section: *An inclusion of this is an admission of its importance*. We can reduce the proportion of abstract nouns and simultaneously increase the proportion of verbs by transforming it as follows: *Since she included this, she is admitting that it was important*. Note that the transformations alter many other stimulus dimensions usually included in advice for readable writing—short clauses, personal pronouns, and word frequency.

It is important to note that in this case the additional explicitness of the verb does not come from its stem—from its content morpheme—but from its function morphemes. Tense, voice, mood, aspect, number, and person are explicitly stated by the inflectional affixes, pronouns, and auxiliary verbs. When we transformed *inclusion* to *included*, tense, voice, aspect, and mood were made specific. More generally, the finite verb is far more explicit than the abstract noun.

Consider the more readable grammatical transformations of our examples above. Note that they contain more personal pronouns than the less readable transformations. Much of the abstractness in scientific writing must be attributed to a reason no more profound than its tradition against "I" and "we" and its tradition against addressing the reader as "you." The scientist who eschews the first two pronouns is simply being modest, but not using them frequently causes him to substitute an abstract noun (*explanation*) for its active verb form (*explain*), to substitute a passive verb for an active verb, to substitute an embedded sentence for a non-embedded sentence.

More generally, if we had more experimental studies of readability such as the ones summarized in Table I, many of the rules for writing readable materials could be precisely stated in terms of grammatical transformations.

D. Packaging Clauses into Higher Units Such as Paragraphs and Stories

There is not much of an experimental nature to say here. Surely most of us believe that the major determiners of readability for adults lie at this level—lie in the associations between clauses and paragraphs, in the overall organization—but psychologists have not yet refined the experimental techniques to investigate this level and linguists are not yet able to describe it.

As far as psychology's share of the problem is concerned, much of it reduces to the problem of developing an efficient, economical measure of comprehension (see subheading D in the next issue). Given an economical measure of comprehension, we can mass produce experiments at this level even though we lack a refined linguistic description. In fact, mass producing enough experiments at this level could provide linguists with the data they need to develop a suprasentence grammar. In the second

half of this paper will discuss developing a set of response measures.

Bibliography

1. Berdiansky, Charles S., "Paired Associate Learning and Response Frequency." University Park: New Mexico State University, 1966. Unpublished masters thesis.
2. Eloomfield, Leonard, *Let's Read*. New York: C. L. Barnhart, Inc., 1963.
3. Byrne, Robert, "Effects of Word Form on Retention," in J. A. Figurel, ed., *Reading as an Intellectual Activity*. New York: Scholastic Magazines, 1963.
4. Carterette, Edward C. and Margaret H. Jones, *Contextual Constraints in the Language of the Child*. Los Angeles: University of California, Los Angeles, 1965.
5. Coleman, E. B. and P. J. Blumenfeld, "Cloze Scores of Nominalizations and Their Grammatical Transformations Using Active Verbs," *Psychological Reports*, 13 (1963) 651-654.
6. Coleman, E. B., "The Comprehensibility of Several Grammatical Transformations," *Journal of Applied Psychology*, 48 (June, 1964) 186-190.
7. Coleman, E. B., "Learning of Prose as Affected by Extra-Experimental Transfer," *Psychological Reports*, 16 (1965) 685-690.
8. Coleman, E. B., "Learning of Prose Written in Four Grammatical Transformations," *Journal of Applied Psychology*, 49 (1966) 332-341.
9. Deese, James, *Structure of Association in Language and Thought*. Baltimore: Johns Hopkins Press, 1965.
10. Desberg, Peter, "Legibility of Whole Word Shapes." Los Angeles: Southwest Regional Laboratory, 1966. Unpublished paper.
11. Gough, P. B., "Grammatical Transformations and Speed of Understanding," *Journal of Verbal Learning and Verbal Behavior*, 4 (1965) 107-111.
12. Himelstein, Margaret D., "Improving the Legibility of Certain Letters." University Park: New Mexico State University, 1965. Unpublished paper.
13. Hockett, C. F., *A Course in Modern Linguistics*. New York: Macmillan Co., 1958.
14. Klare, George R., "The Role of Word Frequency in Readability," *Elementary English*, 45 (January, 1968) 12-22.
15. Martin, Edwin, "Grammatical Factors in Retention," Symposium on Verbal Learning Research and the Technology of Written Instruction, Columbia University, 1966. Unpublished manuscript.
16. Miller, G. A. and Stephen Isard, "Free Recall of Self-Embedded English Sentences," *Information and Control*, 7 (September, 1964) 292-303.
17. Miller, G. R., "Extraexperimental Transfer in Serial Recall," *Journal of Verbal Learning and Verbal Behavior*, 2 (1963) 494-497.
18. Popp, Helen M., "Visual Discrimination of Alphabet Letters," *The Reading Teacher*, 17 (January, 1964) 221-225.
19. Rinsland, Henry D., *A Basic Vocabulary of Elementary School Children*. New York: Macmillan Co., 1945.
20. Rothkopf, Ernest Z., "Learning from Written Sentences: Within Sentence Order in the Acquisition of Name-Clause Equivalences," *Journal of Verbal Learning and Verbal Behavior*, 1964.
21. Templin, M. C., *Certain Language Skills in Children*. Minneapolis: University of Minnesota Press, 1957.
22. Underwood, Benton J. and Rudolph W. Schulz, *Meaningfulness and Verbal Learning*. New York: J. B. Lippincott Co., 1960.
23. Weingartner, H., "Associative Structure and Serial Learning," *Journal of Verbal Learning and Verbal Behavior*, 2 (1963) 476-479.
24. Woodworth, Robert S., "Reading and Eye-Movement," in *Experimental Psychology*. New York: Holt, 1938.

Part II

Measures of Readability and Relevant Populations

2. A Systematic Array of Response Measures

We have seen that reading is a complex hierarchy of subskills, and that materials that are readable according to one subskill may be unreadable according to another. Thus, one of the most important steps in making a systematic study of reading and readability would be developing a sensitive measure for each of the subskills and plotting the R-R functions that relate these measures to one another.

The important point is that different reading subskills are optimally facilitated by different stimulus dimensions. First graders, for instance, will find that some

words are easy to read by whole word memorization and that an entirely different set is easy to read by sounding out. This explains the ease with which an experimenter can "prove" that an extreme teaching system that emphasizes one skill is superior to all others. All he has to do is select a language sample that optimally facilitates his subskill.

If his system depends mostly on whole word learning, for instance, all he has to do is select the most common words in English. Because their response integration is high, a book made up of these words will be very easy to read as long as the child is depending upon whole word recognition. On the other hand, to the extent that the child has to depend upon phonics, the book will be very unreadable because the words are almost all spelled irregularly.

If another experimenter wishes to show that a phonics system is superior, all he has to do is select regularly spelled words that are generated by a minimum of letters (*ill, mill, sill, am, Sam, mam, mass, ass, sass, Sal, Al, lam,* etc.) The child would have to learn only six or seven sound-to-letter associations so these words would be easy to sound out. However, they would be very difficult to learn by whole word recognition because the twenty words would be so similar that it would be difficult to discriminate one from another.

Similar observations can be made for adult reading. Styles that are most readable for light reading may be very inefficient for material that must be more completely mastered.

Clearly we do not want to construct books that optimally facilitate one subskill to the detriment of others. We want to construct books that optimally facilitate a combination of all subskills. At the elementary level, we already possess the techniques for accomplishing this, and they have been described elsewhere (Coleman, 3). That paper described the development of a twenty-word pre-primer that was extremely readable as measured by any of the subskills.

But even before a child finishes reading his first pre-primer, we become concerned with true understanding and must consider the relation of the prose to the universe it describes. Almost everyone—teachers, students, psycholinguists, and even experimental psychologists—are really interested in understanding. We must develop a measure for it, and relate that measure to other measures, such as word-for-word memorization. In carrying out an experimental attack on readability, one of the most important steps is to develop a systematic set of measures at this cognitive level. Let us consider five such measures: (a) answering questions, (b) cloze tests, (c) verification, (d) performance, (e) information gain.

A. Answering Questions.

This is the measure that has been used most frequently to measure a reader's understanding of the passage. The difficulty with this measure is that an investigator can ask easy questions about difficult passages and vice versa. Since the questions that are actually asked are only a small sample from among the ones that could be asked, the questions that one investigator would use might bear little resemblance to those that another investigator would use.

B. Cloze Scores.

The disadvantage of cloze scores is that they do not appear to measure the long range associations in the passage. They are essentially measuring the short range constraints within phrases and clauses and not the more important ones between sentences and paragraphs.

C. Verification.

In several studies, Gough (4) has operationally defined understanding as verification. He showed his reader a picture of an event and a sentence. The reader's task was to say whether or not the sentence truly described the event.

D. Performance.

It is easy to measure the understanding of a passage if the passage assigns the reader a task to perform. The test of understanding is simply to have him perform the task. Coleman (2) used such a measure of understanding in a set of five experiments. His subjects were assigned mathematical tasks. Mathematical tasks are convenient ones since it is easy to generate an infinite number of tests and tasks.

In measuring understanding, we become concerned with the relation between language and events in a real world. The relation will clearly vary with the complexity of the language and also with the complexity of the universe of events that is being described. Since the subjects were assigned mathematical tasks, their universe was a matrix of numbers such as the one in Table I. As their language, they were assigned a miniature language such as the one in Table I.

Thus, we have a miniature language and a miniature universe. We can describe them precisely and we can vary them independently. Despite the restricted vocabulary, the miniature language in Table I has most of the essential features of complete English. It permits one to study most of the variables that affect readability and it permits the experimenters to assign the subject an infinite number of tasks such as *"Sum the even numbers and square the sum."*

The following three experiments used essentially the same method and the same measure of understanding. Sentences that assigned the subject mathematical tasks to perform were prepared for projection (e.g., *Divide the first row's even number by the*

Table I
Miniature Language

Article	*Preposition*	*Pronoun*	*Conjunction*	*Verb*	*Noun*	*Adjective*
a	of	it	and	divide	number	first
an	to		or	subtract	mean	second
the	by			square	row	odd
						even

Matrix

1	3
6	2

second row's odd number). Projection time was the measure, and this was under the control of the subject. He projected the sentence, read it, and kept reading it until he could perform the operation correctly. This measure of total exposure time is sensitive and precise. It can easily be extended to measure suprasentence variables such as overall organization.

Experiment I. The first experiment studied the complexity of the language and the complexity of the universe. This seems a reasonable place to begin since understanding is concerned with the relation between language and events in the universe. Two levels of complexity of the universe (two-by-two matrices and three-by-three matrices) were combined in a factorial design with two levels of complexity for the language. Complexity of the language was increased by adding content words (*multiply, ratio, column, large, small*). The difference between matrices was not significant, but the sentences written in the complex language took far longer to understand (6.9 seconds to 3.1 seconds for twelve-word sentences) and the difference between the two languages increased as sentences became longer.

Experiment II. The second experiment studied the effect of content word ratio. Sentences containing 50 percent content words (*Subtract two from the mean of the rows*) were compared to one that had 75 percent content words (*Subtract two from the first positive odd score*). For twelve-word sentences, the subjects required over twice as much exposure time to understand the 75 percent sentences (10.6 seconds) as they required to understand the 50 percent sentences (5.0 seconds).

Experiment III. A third experiment studied the effect of clause length. Sentences were written that contained the same number of words but varied in clause length; they contained either one or three clauses (e.g., *Square the odd number in the first row. Compute the mean, square it, and add two.*). All thirteen subjects understood the one clause sentence in a shorter mean exposure time (4.2 seconds to 6.3).

This measure of understanding is an economical and sensitive one. With even small numbers of subjects, it yields smooth curves. Clearly, it needs to be extended to tasks other than mathematical ones, such as motor skills to learn, pictures to draw, objects to construct, and so on. It shows some promise of becoming an economical tool for studying the constraints between sentences and paragraphs. Almost all the current descriptors of readability are within-sentence descriptors, but surely the largest determiners of readability are concerned with overall organization, with constraint between paragraphs and larger units.

E. Information Gain.

The above measures are mostly concerned with the gross amount a reader knows *after* he reads a passage. But, as can be measured by cloze tests, the reader always knows something about a passage even before he reads it. To get a measure of information gain, we would subtract this knowledge from the knowledge the reader has about the passage after he reads it.

A little reflection will suggest many uses for such a measure of information gain. In constructing materials for programmed instruction, for example, one is less concerned with ease of reading than with the efficiency with which the material transmits information per unit of time.

Miller and Coleman (5) have suggested one way to obtain a measure of information gain. They required their subjects to guess successively every word in a passage. After

the subject made his guess he was given the correct word as he proceeded through the passage. Thus he saw only the words preceding the one he was trying to guess and not those following it; he was able to use unilateral constraint only. This, of course, is a slight variation of Shannon's (6) well known "guessing game technique." It can give a measure of what the subject knows about the passage before he begins reading it expressed in terms of bits, but Miller and Coleman were content to describe it in terms of percentage of words correctly guessed. After the subject has guessed every word in the passage, he has certainly read it carefully. He was then sent through the passage again in the same fashion, successively guessing each word and being given the correct one as soon as he made his guess. The difference between the proportion of words guessed correctly the first time and the proportion guessed correctly the second time is one measure of information gain. Miller and Coleman called it the IG score, and correlated it with a number of other descriptors.

The measure of information gain is an important one and merits devoting considerable time and expense to its refinement. Among other issues, it points up the distinction between inherent and superfluous complexity. When we try to improve readability, we must distinguish between these two forms of complexity. What we mainly want to eliminate is superfluous complexity. Inherent complexity depends largely upon a passage's cargo of new information, and we must be very selective in reducing that. People are suspicious of readability research because they sense that it is easy to improve "readability" by inserting unimportant fillers or by reducing the technical precision of the vocabulary, *i.e.*, by watering down the content.

Paraphrasing the argument: When we improve readability, we want to reduce superfluous complexity and simultaneously fix the amount of new information (inherent complexity) at an optimal level. According to Miller and Coleman's measure of IG, there appears to be an optimum level of redundancy. For college students, passages written at the difficulty level of fifth-grade readers transmitted a maximum of IG.

Education needs to know how to predict the IG a reader will gain from a passage, and how to write passages so that IG is maximized. Of course, many people have expressed this idea before—usually in carping criticisms of readability research—but always in a vague form that could not lead to quantifiable readability research.

Summary.

One of the most important tasks in mounting an experimental attack on readability is to develop a systematic array of response measures. In particular, we must develop efficient, economical measures of understanding and of information gain. Surely the major determiners of readability lie in the constraints between paragraphs and higher units. However, because our language is so rich, because it offers so many mechanisms for recoding and paraphrasing and transforming what we read, there appears to be no practical way to study these determiners of readability until we develop an efficient measure of understanding.

3. Relevant Populations

Providing the matrix of S-R functions needed to found a technology of readability is a massive undertaking. An economical way to begin would be by replicating previous studies of verbal learning.

The practical application of many of these studies is disguised because their independent variables are described as hy-

pothetical constructs. Once the terminological underbrush is cleared away, however, thousands of these experiments could contribute to our knowledge of readability if they were only replicated on relevant populations—relevant populations of readers and also relevant populations of language materials. For example, most of the experiments of legibility and typographical arrangement have been made using college sophomores as subjects, but these variables are not potent sources of variance for adult readers. They are, however, potent sources for first graders. Perhaps many of the experiments of legibility and related variables should be replicated using first graders as subjects.

The shortcoming that reduces the usefulness of most studies of verbal behavior, however, is an irrelevent sample of language material. The majority of these studies were performed using samples of language materials such as nonsense syllables or disconnected lists of words. A great many of these studies would give us useful insights into readability if they were replicated using relevant samples of language materials.

Generalizing to the Language Population. In experimental studies of readability, the experimental variable is a characteristic of language; it is a stimulus dimension such as word frequency or clause length. The experimenter will use only a small sample of words or clauses—perhaps thirty or forty—but almost inevitably he will claim that his experimental results hold true for language in general. He will *generalize* beyond his sample of language material to a population of materials so let us say that language materials represent a *generalization variable* (in mathematical statistics, the term is *random variable*).

Whenever an experimenter publishes a study investigating a language characteristic, he is implying that his conclusions can be generalized to a language population. But out of the thousands of such studies that have been published, less than a dozen experimenters performed a statistical test that would permit generalization to the language population. The conclusions of many of these studies should be re-examined statistically. The laws of probability guarantee that a certain proportion of them could not be replicated using a different sample of language materials.

The reader should remind himself that a significance test promises that another experimenter could probably replicate the results using a different sample of the generalization variable. Psychology is a science that studies the behavior of organisms, so almost inevitably the generalization variable is represented by the sample of subjects. That is, in psychological experiments, we want to generalize to behavior; we want to be able to say that the experimental results could be replicated if someone else repeated the experiment using a different sample of animals.

In experimental studies of readability and other studies of language behavior, other generalization variables are important. An experimental conclusion about a characteristic of language is useless if it cannot be replicated using a different sample of language materials. Surely no one is interested in an experimental conclusion that holds true for only a dozen or so words or nonsense syllables.

In the thousands of published studies of verbal behavior (with a handful of exceptions), however, the statistical tests only guarantee that the results could be replicated *provided* the experiment were repeated using exactly the same sample of language material. This is important enough to bear paraphrasing in a paragraph of its own.

With less than a dozen exceptions, the significance tests used in most experiments

of verbal behavior do *not* promise that the results could be replicated if another experimenter repeated the experiment using a different sample of language materials. The significance tests promise only that the conclusions can be generalized provided the experiment is repeated using exactly the same set of language materials.

Consider a simple experimental study of readability that compares the difference between active verb sentences and nominalized sentences. Surely no one would be interested in the conclusion of such an experiment if he were told that it had to be restricted to twenty or thirty particular English sentences. Such conclusions would be as useless as conclusions that held true only for a specific twenty or thirty people. In this experiment, language materials have to represent a generalization variable. The experimenter would use only a small sample of sentences, but he would certainly want to generalize beyond them to the population of English sentences.

The purpose of this section is to raise a question, not to provide an answer. This is not the time or place to go into the mathematics of statistical design. It is important, however, to establish that there are statistical designs that allow the experimenter to generalize simultaneously across two generalization variables. They have been described elsewhere (see Coleman, 1, and the more mathematically sophisticated references given in that paper).

It is important to note that the generalization variable represented by language materials should be entered as a formal variable in any experimental study of readability. If the experimenter wishes to generalize beyond the sample of language materials he actually uses in the experiment to a population of such materials, he must draw his sample of materials very carefully, at least as carefully as he draws his sample of subjects. For example, in a study comparing active verb sentences to nominalizations, one way to draw a representative sample of English sentences would be as follows: enter a library, draw a book at random, open it to a page at random, and read until you come to the first nominalized sentence. Draw another book at random, open it to a page at random, and read until you come to the first active verb sentence, and so on. Drawing a representative sample for his generalization variable is actually more important than the mathematics of the significance tests; however, an experimenter studying the characteristics of language should acquaint himself with the experimental designs that allow simultaneous generalization across two samples.

Perhaps, if we omit the mathematical detail, it will be worthwhile to describe at least one experimental design that permits simultaneous generalization across two samples. The simplest experimental design that permits simultaneous generalization simply confounds the two samples into a single one. That is, each subject is given a different set of language materials. In our nominalization-active verb experiment, for example, each subject would read a different set of sentences randomly selected as described above. With some qualifications that are too fine to bother us, the significance test used in an ordinary treatments-by-subjects design would permit simultaneous generalization across both samples.

In closing this section, perhaps we should note explicitly that education is concerned with a number of generalization variables other than language and subjects. Frequently educational researchers wish to generalize their conclusions to populations of teachers, populations of schools, *etc.* Almost inevitably, they have approached this problem by replicating massive, expensive experiments a number of times. Sometimes

these experiments could have been performed at a fraction—a small fraction such as one tenth or one hundredth—of the cost if the generalization variables had been entered as formal variables in the experimental design so that efficient designs (such as confounding the samples) could have been used in a tightly controlled laboratory study.

Surely millions of dollars have been spent comparing the Initial Teaching Alphabet to traditional orthography; however, using a matched pairs design and a tightly controlled laboratory presentation similar to that of Staats, *et al.* (7), the difference between the Initial Teaching Alphabet and conventional spelling can be demonstrated with fifteen matched pairs of subjects. The experimenter only needs to collect an adequate number of responses from each subject each day. Furthermore, the controls of the laboratory situation could eliminate most chances of Hawthorne effects because they would reduce the role of the teacher by programming or by the double blind technique. As a second furthermore, by confounding the sample of subjects with the sample of language materials (each matched pair of subjects reads a different, randomly drawn sample of books), the experiment could be designed to allow simultaneous generalization to the population of language materials as well as the population of learners. As a third furthermore, generalization to a population of teachers could be permitted by formally adding this to the experimental design as another generalization variable.

Conclusions

The major purpose of my two papers was to advocate a program of experimental studies of readability. To make massive improvements in the readability and teaching capacity of textbooks at the adult level, we must first develop measures of comprehension and information gain. We must develop techniques and instruments for mass producing experiments at this cognitive level.

The techniques for making such massive improvements in elementary reading materials, however, are almost within our grasp. With only a trifling investment, we can construct books that are so readable that they will completely change the notion of reading readiness. Almost all the fundamental concepts of reading readiness—the concept that shapes represent sounds, the habit of left-right progression, *etc.*—can be taught as the child actually reads little books.

Let us summarize the research strategy that will lay the data base necessary for constructing such books. First we will analyze beginning reading material into its most basic units, letters and combinations of letters. Then we will measure the ease with which the children learn to make all the different reading responses to each element. Once the letters and letter combinations are thus calibrated and rank-ordered as to difficulty, it will be possible to prepare reading materials that are far easier than those currently available. In brief, the beginning vocabulary will be restricted to a few of the easier words—words that are composed of a few of the easier letters.

By using cartoons, we have prepared books so simple that preschoolers have learned to read them after a single fifteen-minute teaching lesson. As Figure 1 shows, the books are not deathless literature. They are mostly pictures—only one, two, or three different words—only two or three different letters. These books, however, simple as they are, are quite effective in teaching two of the most basic concepts in reading—the concepts that the shapes represent sounds and the habit of left-right progression.

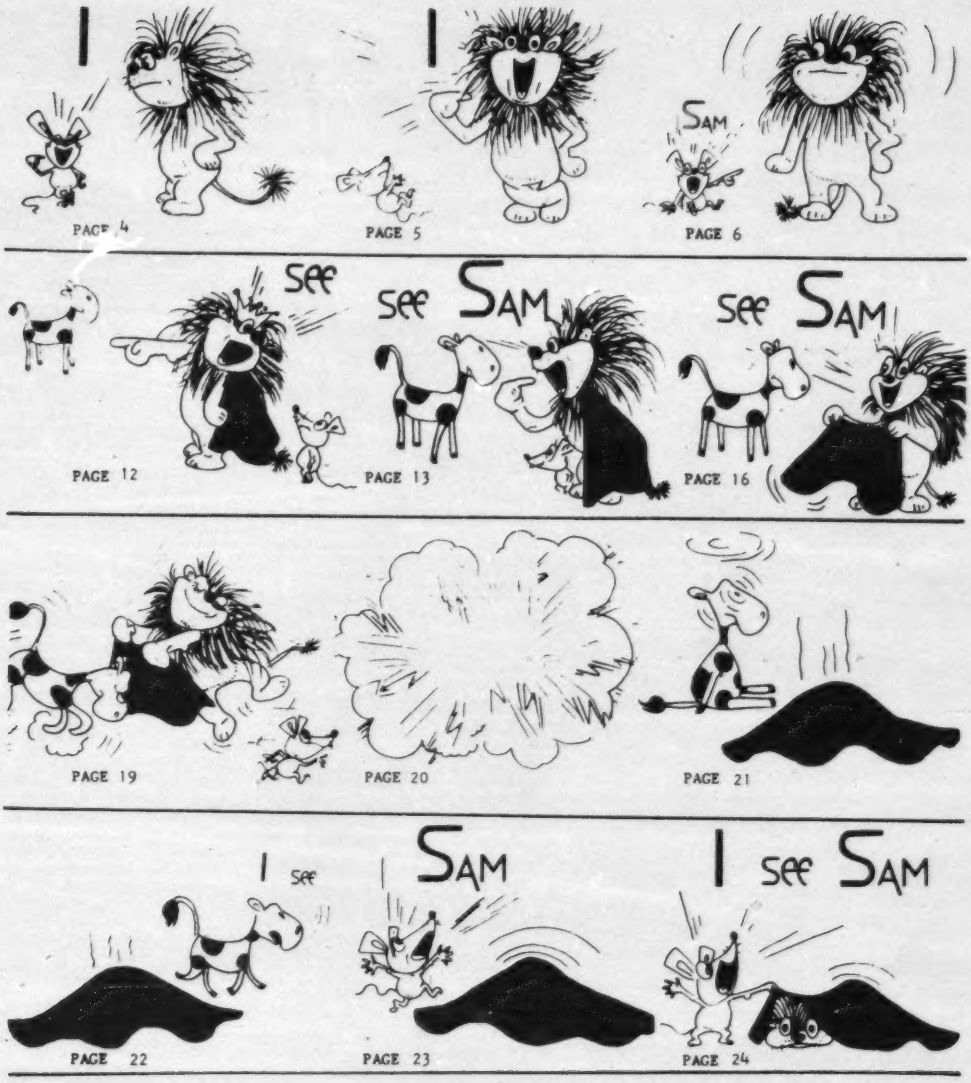

Figure I

Excerpts from a book that most five year olds can learn to read in a 20-minute lesson. In its last test, over 75 percent of a class of kindergarteners learned to read it after a single lesson.

Later books add a few more letters and words and teach other basic concepts: that words can be analyzed into letters and sounds, that letters and sounds can be reintegrated into words, and so on. In short, these books teach the concepts of reading, spelling, and printing simultaneously; they teach them by having the child read enjoyable little books rather than by memorizing long lists of words, letters, phonic concepts, and spelling irregularities. A readable book is an efficient teaching machine. If we teach the child the basic concepts, and give him books that he en-

joys reading, much of his problem is solved. He begins reading and teaching himself.

Once we collect the systematic array of S-R functions advocated in this paper, it will be possible to write little books that are readable by preschoolers after the sort of casual instruction that could be given by Captain Kangaroo on television. Before they enter kindergarten, children could have read little cartoon books that would have made them accomplished masters of the basic concepts of the written communication skills.

An example of such a cartoon book is given in Figure I. It is a simple matter to teach a four year old to read a book such as this. He need only distinguish three shapes (*I, Sam, see*), and this is within the mental competence of a bright chimpanzee.

An average five year old can learn to read such a book in a single 20-minute lesson. This book and its lesson have been tested in several kindergarten groups and have always had over 65 percent of the children reading after the single lesson.

The book and the lesson are in a continual state of refinement. In its last test (Larch Elementary School, Lennox, California), the most refined version (the lesson resembled an animated cartoon) had all but two of 22 children reading after the single lesson. In an individually administered test on a different book the next day, only four of the children missed more than half the book's 19 words.

In short, by manipulating stimulus dimensions that affect readability, we can create books as readable as Figure I. Then, by exploiting the child's love affairs with television and with animated cartoons, we can teach him to read immediately. We can extend Operation Head Start into Operation Running Start—a running start for all our children.

Bibliography

1. Coleman, E. B., "Generalizing to a Language Population," *Psychological Reports*, 14 (1964 219-226.
2. Coleman, E. B., "Developing a Technology of Written Instruction: Some Determiners of the Complexity of Prose," in Ernst Z. Rothkopf (Ed.), *Symposium on Verbal Learning Research and the Technology of Written Instruction*. Rand McNally, in press.
3. Coleman, E. B., "Experimental Studies of Learnability: Their Contributions to a Technology of Written Instruction," IRA Preconvention Institute III, Research Designs in Reading, 1966.
4. Gough, P. B., "Grammatical Transformations and Speed of Understanding," *Journal of Verbal Learning and Verbal Behavior*, 4 (1965) 107-111.
5. Miller, G. R. and E. B. Coleman, "A Measure of Information Gained During Prose Learning," *Reading Research Quarterly*, in press.
6. Shannon, C. E., "Prediction and Entropy of Printed English," *Bell System Technical Journal*, 30 (1951), 50-64.
7. Staats, A. W., J. R. Finley, K. A. Minke, and M. Wolfe, "Reinforcement Variables in the Control of Unit Reading Responses," *Journal of Experimental Analysis of Behavior*, 7 (1964) 139-149.

JOHN R. BORMUTH

The Cloze Readability Procedure*

Educators must have a valid method of finding out whether instructional materials are understandable to their students, for students must acquire much of their knowledge by reading written instructional materials. The conceptual or aesthetic merits of a set of materials is a primary consideration in the selection of materials. But materials can have little educational value if they are written in language that is so complex and obscure children cannot understand the contents. The readability formulas presently available are too inaccurate to provide educators with much help in this matter. And, while it is now possible to construct highly valid readability formulas (6), such formulas are still in the development stages.

This paper will describe ways in which teachers can use the cloze readability procedure to determine if instructional materials are understandable to children. A somewhat oversimplified description of the cloze readability procedure includes these steps: (a) passages are selected from the material whose difficulty is being evaluated, (b) every fifth word in the passages is deleted and replaced by underlined blanks of a standard length, (c) the tests are duplicated and given, without time limits, to students who have *not* read the passages from which the tests were made, (d) the students are instructed to write in each blank the word they think was deleted, (e) responses are scored correct when they exactly match (disregarding minor misspellings) the words deleted. When the tests have been made properly, a student's score can be interpreted as a measure of how well he understands the materials from which the tests were made.

The first section of this paper will discuss the validity of the cloze readability procedure. The second section will describe how the cloze readability procedure may be used to place a child in materials that are graded in difficulty, and to select materials that are understandable for a group.

Validity of the Cloze Readability Procedure

Logical or Face Validity

Similarity to Conventional Tests: At first glance a cloze readability test appears to be a completely different kind of test and some authors have made much of this fact, attempting to construct all sorts of mystical theories about cloze tests. Some have professed to see a similarity between the processes involved in responding to a cloze test and the clozure phenomenon observed in the perception of geometric figures. Indeed, it was from just this kind of conjecture that cloze procedure got its name.

On closer inspection it can be seen that many of the items in cloze readability tests are identical to those found in reading

Dr. Bormuth is an Associate Professor of Education at the University of Chicago. The preparation of this paper was accomplished with the support of a grant from the Center for the Study of the Evaluation of Instructional Programs at the University of California, Los Angeles.

* This article has been reprinted from the April, 1968, issue of *Elementary English*, Volume XLV, pages 429-436.

comprehension tests made by conventional methods and that the processes required to fill cloze blanks are probably not different from those required to answer conventionally made items.

Conventional completion test items are made by simply deleting a word or phrase from a sentence. For example, given the sentence *The boys rode horses*, it is possible to make the completion questions,- - - -*rode horses*, *The boys- - - -*, *The boys - - - -horses*, and *The boys rode- - -* . The familiar wh- questions are made in much the same manner. A word or phrase is deleted, a wh- phrase is inserted in its place, and the sentence is transformed so that it begins with the wh- phrase. This gives the questions. *Who rode horses? What did the boys do? What did the boys do to the horses?* and *What did the boys ride?* As in cloze tests, the correct answers to these questions are the words or phrases deleted.

Contrasts with Conventional Tests: But items made by cloze and conventional test making procedures differ in several important respects. First, in a cloze readability test, only one word is deleted at a time while in conventional tests, whole phrases and clauses may also be deleted. Further, in cloze readability tests, structural words may be deleted. (Structural words consist of classes such as articles, prepositions, conjunctions, modal and auxiliary verbs, *etc.*) But in conventional tests, only lexical words (consisting roughly of verbs, nouns, adjectives, and adverbs) may be deleted by themselves.

A second major difference is the fact that cloze readability tests are made only from the sentences in the text while conventional tests items may be made either from sentences in the text or from the sentences that can be derived from the text (5). A sentence may be derived by any, or all, of three processes. The simplest type of derived sentence is obtained by transforming the sentence. For example, the sentence *The boys rode the horses* can be transformed into *Horses were ridden by the boys* and then transformed again into questions such as *By whom were the horses ridden?* Sentences can also be derived by substituting synonymous words or phrases for the words or phrases in the sentence in the text. Finally, derived sentences may be obtained by explicating the statements implied by the fact that two sentences are contiguous. Consider the sentence, *The boys got home first* followed by *They rode horses.* The contiguity of these sentences implies the sentence *Riding the horses caused the boys to get home first.* Only conventional test items can be made from sentences derived from but not actually in a text.

The third major contrast is the fact that cloze readability tests are taken by students who have not read the undeleted version of the passage.

Probably too much has been made of these contrasts between cloze and conventional tests. The student has eighty percent of the text on which to base his responses, so his responses very much depend on his ability to understand the text. Also, the fact that he has not read the original text may require that he uses processes similar to those required to answer questions made from derived sentences plus a sensitivity to the author's style and the tone of the passage. However, the contrasts do exist and so the question must be referred to the researcher, the final arbiter of such disputes.

Concurrent Validity

There have been many studies in which the objectives were to determine if scores on cloze readability tests correlate with scores on conventionally made comprehension tests and if the cloze difficulties of

passages correlate with the difficulties of the passages as determined by conventional tests. By now, this literature has grown too large to review in detail in this paper. Since extensive analyses of this research are available in Rankin (13) and Bormuth (5) only a few of the most decisive studies will be mentioned here. In general, these studies seem to show that cloze and conventional tests measure the same processes.

Validity as a Measure of Comprehension Ability: Cloze tests seem to measure a wide variety of comprehension responses. Wilson Taylor (18), the originator of the cloze procedure, found a correlation of .76 between scores on a cloze readability test and scores on a multiple choice test made from the same passage. Bormuth (3) constructed tests to measure the comprehension of vocabulary, explicitly stated facts, sequences of events, inferences, causal relationships, main idea, and author's motive in each of nine passages. He gave these and the cloze readability tests to elementary school students. The correlations between the cloze and conventional tests over each passage ranged from .73 to .84. When the correlations were corrected for the unreliabilities of the tests, the correlations approached 1.00. When the conventional tests were scored to obtain scores for each type of item, and then factor analyzed along with the cloze test scores, only one factor emerged. Kohler (11) performed a somewhat similar factor analysis study using tenth-grade students and obtained similar results.

To determine if cloze readability tests measure some of the subtler responses to language, Bormuth and Hook (9) studied the correlations between cloze readability tests and tests of ability to recognize an author's style. Two tests of ability to recognize authors' styles were made by selecting short passages from several of the works by an author, works that the students in the experiment had not read. The passages were then mixed with passages from works by other authors and given to college students who had just studied two books by each of the authors. Scores based on ability to correctly identify the passages by an author had significant correlations with scores on cloze tests made from passages taken from works of the author.

Validity as a Measure of Comprehension Difficulty: The mean percent of items a group answers correctly on a cloze readability test seems to provide an accurate measure of the difficulty of the passage, almost regardless of how difficulty is measured. Sukeyori (16) found a correlation of .83 between the cloze readabilities of passages and the combined subjective ratings of the passages by three judges. Bormuth (8) found a correlation of .92 between the cloze readabilities of passages and the difficulties of the same passages as measured by multiple choice tests. Subsequently, Bormuth (2) used each of the four forms of the Gray Oral Reading Paragraphs. He found correlations ranging from .90 to .95 between the cloze and word recognition difficulties of the paragraphs and correlations ranging from .91 to .96 between the cloze and the comprehension difficulties of the paragraphs.

Experimental Validity

One study (8) approached the problem of cloze readability test validity experimentally. Passages were varied systematically in language complexity and subject matter. Both cloze and multiple choice tests were made over each passage and given to students at three different grade levels. Each factor and the interaction between language complexity and subject matter produced significant effects on both kinds of test scores. The effects were ap-

proximately proportional on the two kinds of test scores.

Validity of the Procedure Itself

A cloze test can be made, administered, scored, and interpreted in a number of ways. The particular set of alternative procedures actually used in the cloze readability procedure represent what researchers have found are the best procedures when validity, economy, and convenience are considered simultaneously.

Test Construction: The practice of deleting every fifth word is followed because it is simple and economical to use and because it provides the greatest number of items possible for a given passage and thereby provides the most reliable measure of passage difficulty. While deletions may be less frequent than every fifth word, MacGinitie (12) has shown that when a deletion system leaves less than four words of context between items, a student's ability to answer an item begins to depend heavily upon whether he was able to answer correctly the adjacent items. When this occurs, the scores become difficult to interpret in any meaningful way.

Administration: Cloze tests may be administered either with or without the student reading the passage from which the test was made. Research shows that the two methods are about equally valid. Taylor (17) found that scores on cloze tests administered after students had read the passage had slightly higher correlations with scores on comprehension tests. Rankin's (14) studies showed the same results. But the data in both studies showed that this effect probably was the result of scores being somewhat more variable than when the students had not read the passage, an effect that is more economically and easily obtained by simply adding a few items to a test. Therefore, because of the savings in testing time and preparation of materials and because results which are just as valid can be obtained, it is the most desirable procedure to give the tests to students who have not read the passages from which the tests are made.

Scoring: A student's response can differ from the deleted word in semantic meaning, grammatical inflection, and spelling. Users of the cloze readability procedure have settled on the practice of scoring correct just those responses which exactly match the deleted word, but minor misspellings are disregarded. This practice is based on findings by Taylor (18) Rankin (14) and Ruddell (15) that including synonyms as correct responses slightly increases the correlations between cloze scores and scores on comprehension tests. But their data show that it does so simply by increasing the variability of the scores, an effect that is far more easily obtained by adding a few items to the test.

Bormuth (7) classified responses into three categories depending on whether root forms of the responses were identical to, synonymous with, or semantically unrelated to the word deleted and then further classified the responses in each category according to whether they were grammatically correct or not. Only the grammatically correct responses had significant correlations with a comprehension test. When the scores based on the responses exactly matching the deleted words were held constant, all the other correlations dropped to zero. Hence, responses that exactly match the deleted words furnish the most valid measures of comprehension.

Interpretation: There is some value in knowing that one passage is more difficult for students than another. But a cloze readability score has little value unless a teacher can say that the score does or does not represent a satisfactory level of perform-

ance on the materials from which the test was made. A standard has long been accepted (19) for conventional comprehension tests and this standard is widely used in practice (1;10). It asserts that materials are suitable for use in a child's instruction when he is able to answer correctly 75 percent of the questions asked him about the materials. The materials are said to be suitable for his independent study when he can answer 90 percent of the items. Bormuth (2;4) found that a score of 75 percent on conventional comprehension tests is comparable to a score of 44 percent on a cloze readability test made from the same passage and that answering 57 percent of the cloze items is comparable to answering 90 percent of the items on conventional comprehension tests.

Summary

In general, the studies of the validity of tests made by the cloze procedure seem to justify four assertions. First, cloze readability tests provide a valid measure of a student's reading comprehension ability. Second, the cloze readability procedure provides a valid method of measuring the comprehension difficulties of passages. Third, the procedure itself seems to incorporate both the most valid and the most economical of the possible alternatives for designing a cloze readability procedure. Finally, cloze readability scores can be used to judge the suitability of materials.

Applications

Perhaps the most important advantage of the cloze readability procedure is that it requires little training in testing technology of the person who wishes to use the procedure in many of its important applications in schools. This section will describe two of the most important types of applications.

Placing Students in Graded Materials

One of the most critical tasks a teacher must perform in reading instruction is to place the students in basal or other practice reading materials that have the appropriate level of difficulty. The materials must present enough difficulty to permit practice of reading skills but not so much difficulty that the student develops faulty reading habits and feelings of frustration. The method of testing materials that is presently advocated calls for the teacher to have the student read orally a sample passage from each level of materials and then answer questions about what he just read. The teacher notes the adequacy of both the child's word recognition and comprehension. This method has undeniable advantages, but because it requires much time to administer and a relatively high degree of training on the part of the teacher, it is seldom used.

The cloze readability procedure can be used to make a set of tests which permit the rapid and accurate assignment of students to materials having an appropriate level of difficulty. The basic steps of this procedure consist of making a cloze test over each of the texts in the series, administering them to the students, scoring the tests, and placing the students in the materials on which their cloze readability score fell into the appropriate range.

The most critical part of this procedure is to select a test that best represents the difficulty of a text. This is done by initially making several tests from passages randomly selected from the book, finding the mean difficulty of the entire set of tests, and then discarding all but the test having a difficulty closest to the mean difficulty of all the tests.

A step by step description of this process includes these steps. (a) Select several passages from a book, say from six to twelve, using some random method for selecting

the passages. Each passage should begin at the normal beginning of a paragraph and each should be at least 250 words in length. The test, for reasons of both convenience and test reliability, should contain exactly 50 items. (b) The tests are then given to students at the grade level at which the text is most commonly used, 25 to 30 students are usually enough to obtain reliable results. (c) The mean score on each test is calculated and then the mean of the mean scores is calculated. (d) Finally, the test whose mean score is closest to the mean of the entire set of tests is selected and the rest of the tests are discarded. (e) When a test has been selected for each of the texts a teacher is likely to use, the tests can be mimeographed and compiled into booklets which can be administered as group tests. When a student's score falls between 44 and 57 percent on one of these tests, the materials are at the level of difficulty thought to be suitable for use in his supervised instruction. Materials on which a student's score is above 57 percent are suitable for use in his independent study.

The reliability of this procedure depends upon four factors. First, if longer tests are used, the students' scores will be more accurate, but the tests will also require more time to administer. Second, if a larger number of tests are made when selecting the test to represent the materials, the test selected will more accurately reflect the difficulty of the materials. If the tests are to be used frequently, it is probably worthwhile to select the test from a larger number of tests. Third, some materials vary unevenly in difficulty as they proceed. These materials are poorly constructed and their use should, if possible, be avoided. Fourth, the procedures outlined for constructing, administering, and scoring the cloze tests must be followed exactly or the results will be meaningless.

Selecting Materials for a Group

When materials are being considered for adoption by a school, one of the important questions that must be considered is whether the materials are understandable to students. While the quality of the conceptual content of materials is of paramount importance, that content has little likelihood of being learned unless it is presented in an understandable manner. The cloze readability procedure can be used for evaluating the suitability of materials for a group.

Criteria for Evaluation: When the same text is to be used with all the students in a group, two criteria must be used in judging the suitability of the text. First, the materials whose difficulty is most appropriate are those on which the largest number of students can demonstrate a satisfactory level of performance. Second, the best materials are those in which the level of difficulty is fairly uniform throughout the materials. Either all sections of the book should exhibit about the same level of difficulty or the sections should gradually become more difficult as the student proceeds through the book.

Designing the Tests: From the point of view of accuracy in evaluating materials, it would be ideal if a cloze readability test were made over the entire book. But practical necessity demands that testing time and materials preparation be kept at a minimum. Consequently, a procedure for sampling passages is usually required. If done carefully, sampling techniques will lead to accurate results.

The best plan is to divide the book into sections, select two or more sample passages from each section, and use those passages to make the cloze tests. The passages should be selected by some random process. For example, the paragraphs in a

section might be numbered, the numbers written on slips of paper, the slips shuffled, and two or more slips drawn with the restriction that no two passages should contain the same material. The passages may be as short as a paragraph or as long as seems to make a convenient test. (A 250 to 300 word test fits nicely on one sheet of paper.) But each sample should be a length of continuous text, and it should begin at the beginning of a paragraph. The samples need not be of exactly identical length since the evaluator will be working with percentage scores.

Selecting the Students: The students to whom the tests are given should be either the entire population who will use the book or a sample of students who are truly representative of that population. A good sample of students can be obtained by randomly drawing one or more students from every classroom in which the book will be used or by numbering the students in the population and randomly drawing the required number of students.

Administering the Tests: For practical reasons, it is seldom desirable to give all the tests to every student. It is perfectly permissible to break the group up into subgroups and administer a fraction of the tests to each subgroup. But the selection of subgroups should be done by a random method to avoid biasing the means of the individual tests.

Analyzing the Results: The analysis of students' scores begins by converting each score on each passage separately into a percentage score. A useful analysis is to determine what proportion of the scores were 44 percent or larger. This provides a measure of the proportion of students from whom the book, as a whole, is or is not suitable. Another analysis is to calculate this proportion of scores but to calculate it separately for each passage and for each section. When a frequency distribution is made of these proportions or when their standard deviation is calculated, the evaluator obtains a measure of how variable the different sections of the book are in difficulty. The evaluator can also plot a graph of the section difficulties to determine how the difficulty changes across sections, but for this purpose it is best to calculate the means and plot them.

Comparing Alternative Texts: If the evaluator has some training in statistics or has access to a consultant who is competent in these matters, it is possible to compare texts and even compare two or more textbook series. Using an analysis of variance and a carefully designed procedure for selecting the sample passages from each text or series it is possible to find out if texts differ in difficulty, if the texts differ in the variability of their difficulty, and if the texts increase systematically in difficulty as they proceed from section to section and text to text.

Assigning a Grade Placement to a Text: Often it is not possible to administer the cloze readability test to a representative sample of the students who will use it. And even if it is possible to do so within one school system, the results are seldom generalizable to other school systems. Consequently, it may be desirable to express the text's difficulty in a more generally useful form, as a grade placement. When a grade placement number is attached to a book, it can be interpreted as the average level of reading achievement attained by children who are able to read the book at the minimum level of comprehension.

The method is fairly simple, but it requires having a reading achievement grade placement score on each student and it

requires some amount of calculation. The grade placement score is first correlated with the student's cloze readability percentage score. Then, using the common regression prediction formula, the evaluator calculates the reading grade placement score that is comparable to a score of 44 percent on the cloze readability test. This is the number assigned as the grade placement of the material. It can be calculated either for each passage in the text or for the text as a whole. The grade placement number is then useful to teachers in other schools, who, if they use the same reading achievement test, can calculate the proportion of their students who will be able to understand the text at a minimum level of competence by counting the number of students who have reading achievement scores falling above the grade placement number assigned to the text.

Concluding Remarks

Educators have long felt the need to adapt instruction to the individual differences among their students but their efforts to do so have often been hampered by the absence of practical procedures for attaining this objective. An important way to adjust instruction is to place in each child's hands materials having a level of difficulty that is appropriate to his level of reading skill. The cloze readability procedure provides several devices to help accomplish this objective.

Bibliography

1. Betts, E. A., *Foundations of Reading Instruction*. New York: American Book Company, 1946.
2. Bormuth, J. R., "Cloze Test Readability: Criterion Reference Scores," *Journal of Educational Measurement*, (1967). (in press)
3. Bormuth, J. R., "Factor Validity of Cloze Tests as Measures of Reading Comprehension Ability." Paper read at the American Educational Research Association in February, 1967.
4. Bormuth, J. R., "Comparable Cloze and Multiple Choice Comprehension Test Scores," *Journal of Reading*, 10 (February, 1967) 291-299.
5. Bormuth, J. R., *The Implications and Use of Cloze Procedure in the Evaluation of Instructional Programs*. Center for the Study of Evaluation of Instructional Programs, University of California, Los Angeles, Report No. 3, 1967.
6. Bormuth, J. R., "Readability: A New Approach," *Reading Research Quarterly*, 1 (1966) 79-132.
7. Bormuth, J. R., "Validities of Grammatical and Semantic Classifications of Cloze Test Scores," *Proceedings of the International Reading Association*, 10 (1965) 283-286.
8. Bormuth, J. R., *Cloze Tests as Measures of Readability*. Unpublished doctoral dissertation, Indiana University, 1962.
9. Bormuth, J. R. and Ora M. Hook, "Cloze Tests as a Measure of Ability to Detect an Author's Style," *Proceedings of the International Reading Association*, 10 (1965) 287-290.
10. Harris, A. J., *Effective Teaching of Reading*. New York: David McKay Company, 1962.
11. Kohler, E. T., "A Factor Structure Study of Cloze Tests." Paper read at the annual meeting of the National Council on Measurement in Education, New York, 1967.
12. MacGinitie, W. H., "Contextual Constraint in English Prose Paragraphs," *Journal of Psychology*, 51 (1961) 121-130.
13. Rankin, E. J., Jr., "Cloze Procedure—A Survey of Research," *Yearbook of the South West Reading Conference*, 14 (1965) 133-148.
14. Rankin, E. J., Jr., *An Evaluation of the Cloze Procedure as a Technique for Measuring Reading Comprehension*. Unpublished doctoral dissertation, University of Michigan, 1957.
15. Ruddell, R. B., *An Investigation of the Effect of the Similarity of Oral and Written Patterns of Language Structure on Reading Comprehension*. Unpublished doctoral dissertation, Indiana University, 1963.
16. Sukeyori, S., "A Study of Readability Measurement—Application of Cloze Procedure to Japanese Language," (English Abstract), *Japanese Journal of Psychology*, 28 (1957) 135.
17. Taylor, W. L., "Recent Developments in the Use of 'Cloze Procedure,'" *Journalism Quarterly*, 33 (1956) 42-48.
18. Taylor, W. L., "'Cloze Procedure': A New Tool for Measuring Readability," *Journalism Quarterly*, 30 (1953) 415-433.
19. Thorndike, E. L., "Reading and Reasoning: A Study of Mistakes in Paragraph Reading," *Journal of Educational Psychology*, 8 (1917) 323-332.